MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection

Jiaqi Zheng[1], **Xi Zhang**[1*], **Sanchuan Guo**[1], **Quan Wang**[1],
Wenyu Zang[2] and **Yongdong Zhang**[3,1]
[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
[2]China Electronics Corporation
[3]University of Science and Technology of China
{zjq1, zhangx, guosc, wangquan}@bupt.edu.cn, wenyuzang@sina.com, zhyd73@ustc.edu.cn

IJCAI'22 (International Joint Conference on Artificial Intelligence)

220825 Chia-Chun Ho

Outline
of MFAN

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction

Fake News Detection

- With the rapid development of social media, rumors cam quickly spread over these platforms.
  - It lead to significant negative impacts on society.
    - The rumor blaming 5G for the coronavirus pandemic had led to arson attacks on more than 70 cell phone towers in the UK in 2020.
- Due to the large amounts of user-generated content every day.
  - It's desirable to automatically identify rumors to minimize the harmful impacts.

Introduction

Existing Approaches

- Traditional rumor detection models mainly rely on extracting textual features.
  - Either with traditional learning models such as decision trees or DNN-based models such as RNNs & CNN.
- With the prevalent of multimedia, spreaders utilize visual content together with textual content to attract more attention and get rapid dissemination.
  - Fuse textual and visual features based on DNN to produce multimodal post representations, which have shown better performance than solely using the textual data.
- However, one common limitation of these studies is that they didn't consider the graphical social contexts simultaneously.

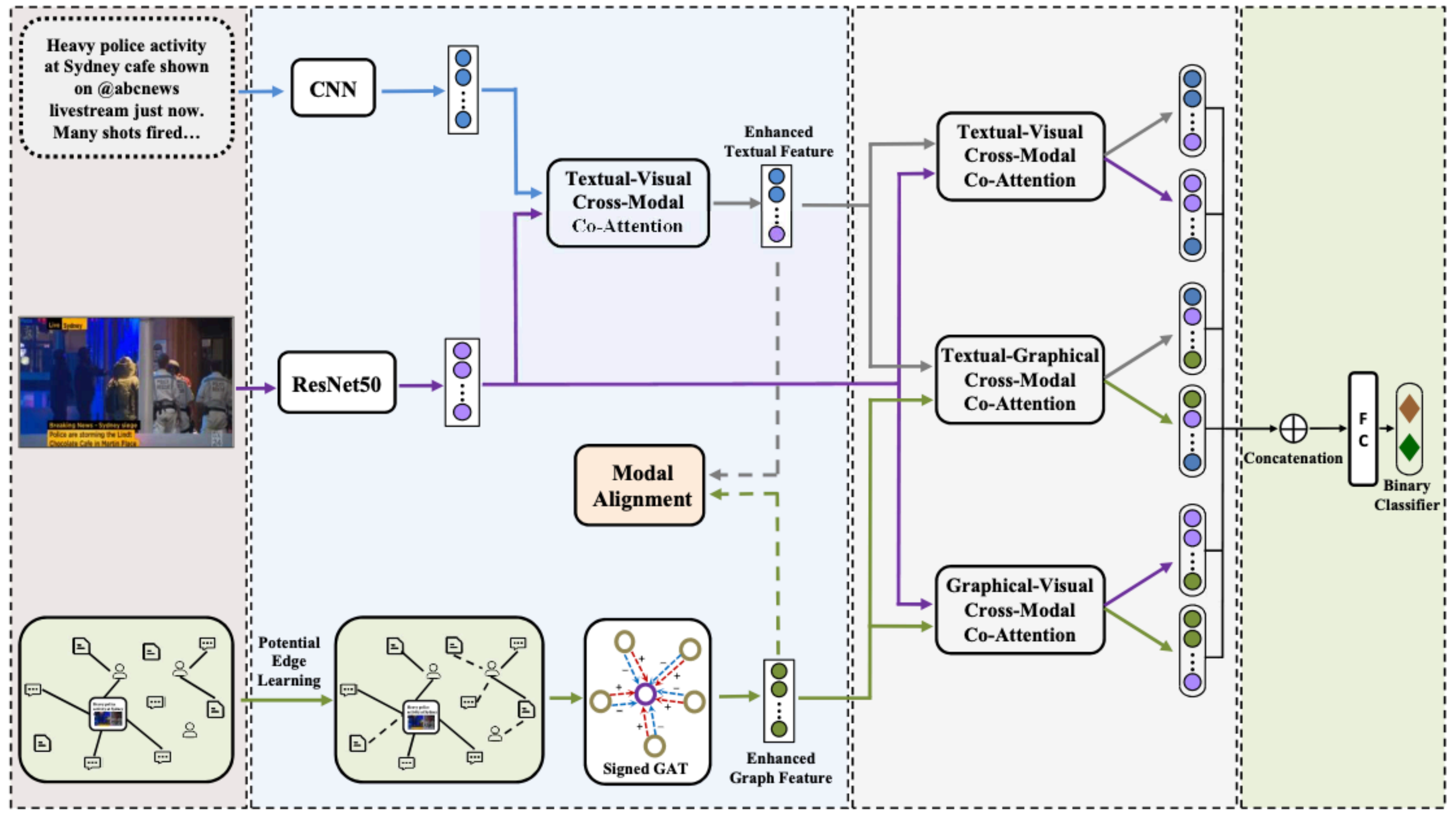
Introduction

Limitations (1/2)

- The existing graph-based detectors suffer from several limitations:
  - The quality of node representation learning depends highly on reliable links.
    - Due to the privacy issue or data crawling constraint, the available social graph data is very likely to lack some important links among entities.
    - Therefore, it is necessary to complement latent links on the social graph to achieve a more accurate detection.

Introduction
Limitations (2/2)

- The existing graph-based detectors suffer from several limitations:
  - There may be various latent relations between adjacent nodes on a graph.
    - While the conventional neighborhood aggregation procedure of GNN may not be able to differentiate their effects on the representation of a target node, leading to inferior performance.
  - How to effectively integrate the learned social graph features with other modality features is less explored in existing studies.



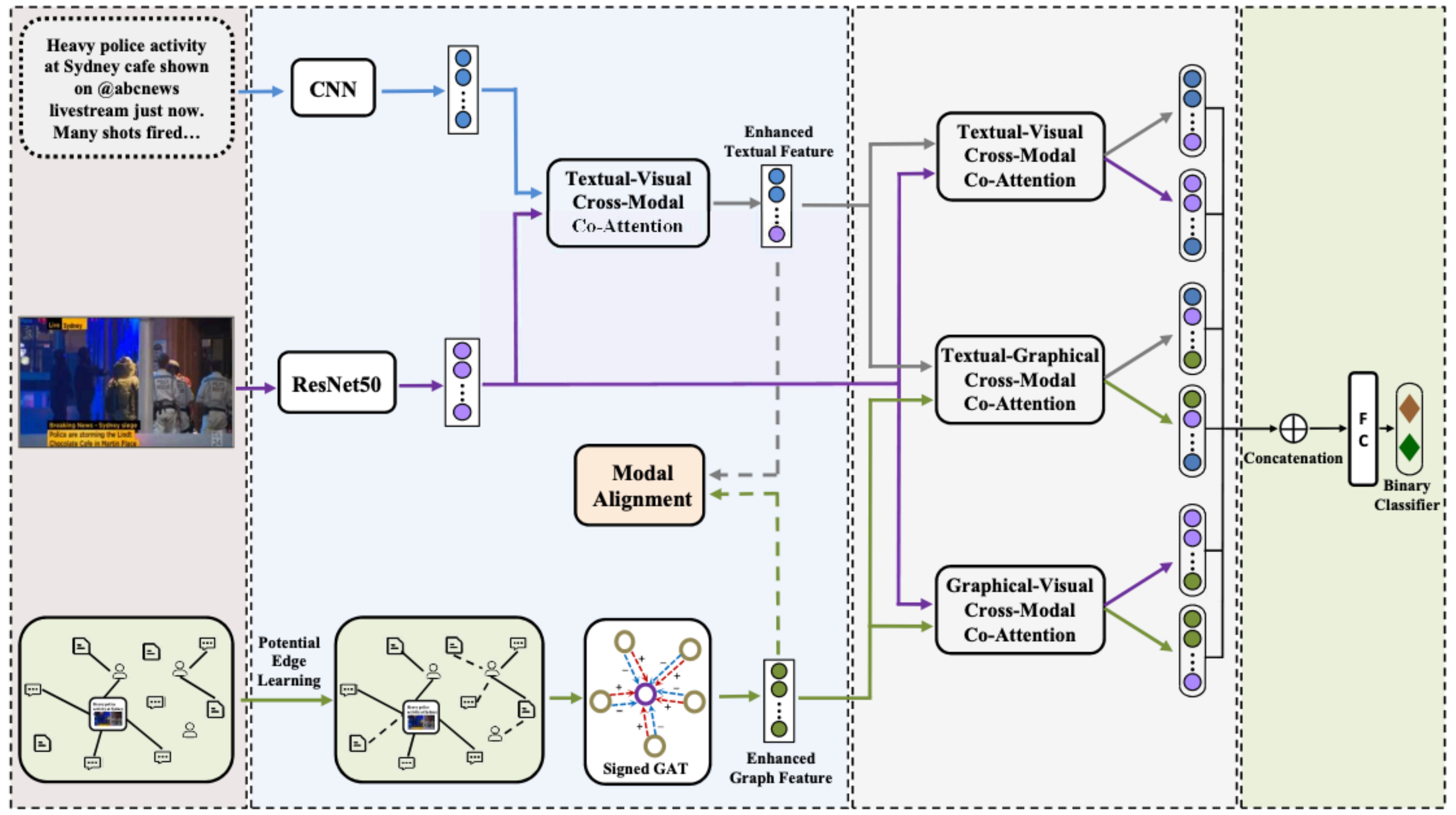
Introduction

MFAN

- Propose the Multi-modal Feature-enhanced Attention Network for FND.
  - First attempt to jointly model textual, visual and social graph features in one framework.
  - Improve the multi-modal fusing mechanism by considering the cross-modal semantic alignment.
    - Specifically, a self-supervised loss is introduced to align the source post representations learned from two distinct views. (textual-visual, social graph view)
  - On the one hand, propose to infer potential links between nodes in the social graph to alleviate the incomplete link issue.
    - On the other hand, utilize a signed attention mechanism to capture both positive and negative neighborhood correlations to achieve better node representations.

Introduction

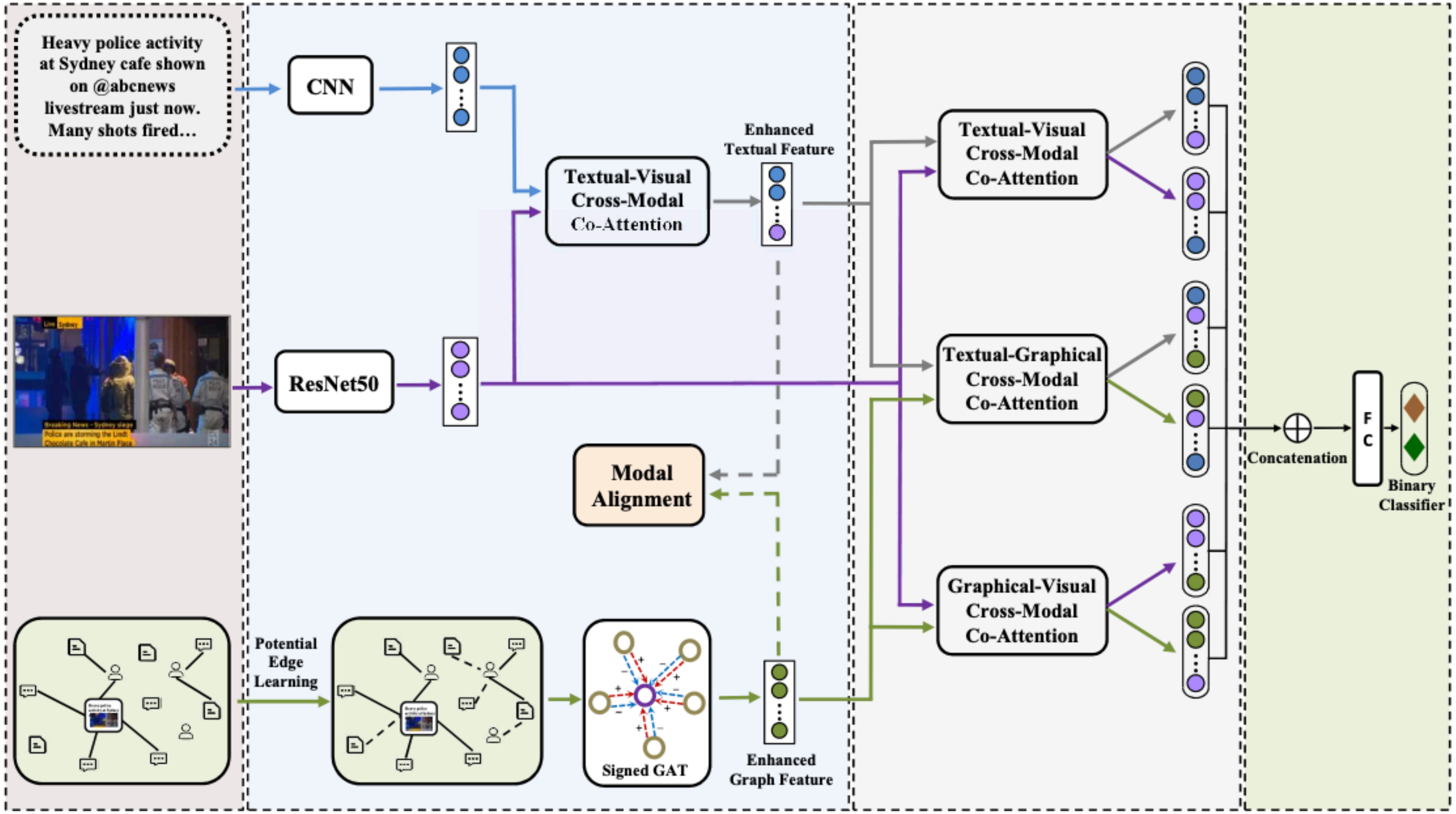
Contributions



- Propose a multi-modal feature-enhanced attention network for FND.
  - Effectively combine textual, visual, and social graph features in one unified framework.
- Introduce a self-supervised loss to align the source post representations in different views to achieve better multi-modal fusion.
- Improve the social graph feature learning by enhancing both the graph topology and neighborhood aggregation procedure.

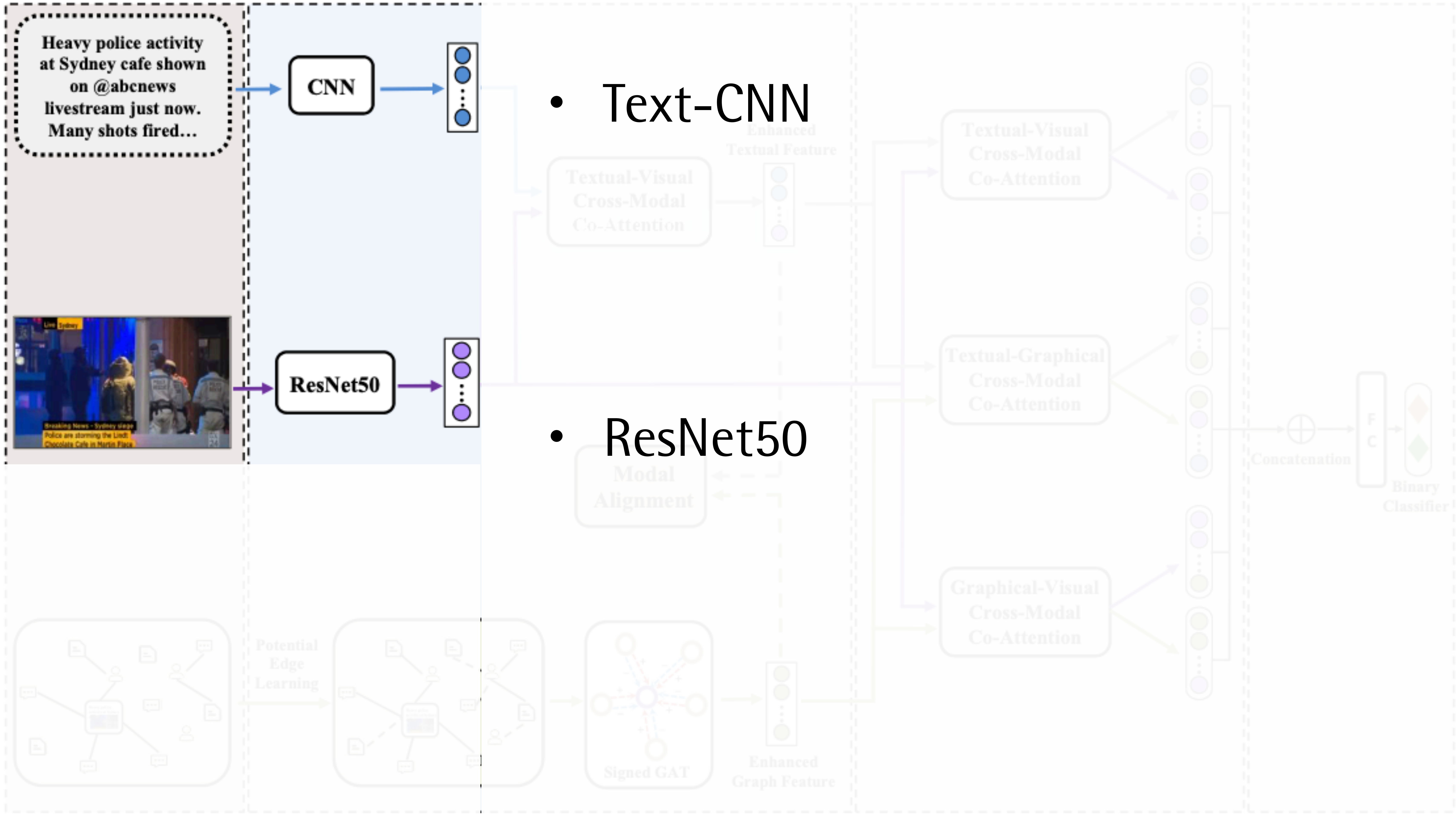
Methodology

Multi-modal Feature-enhanced Attention Networks (MFAN)



Methodology

Textual & Visual Feature Extractor

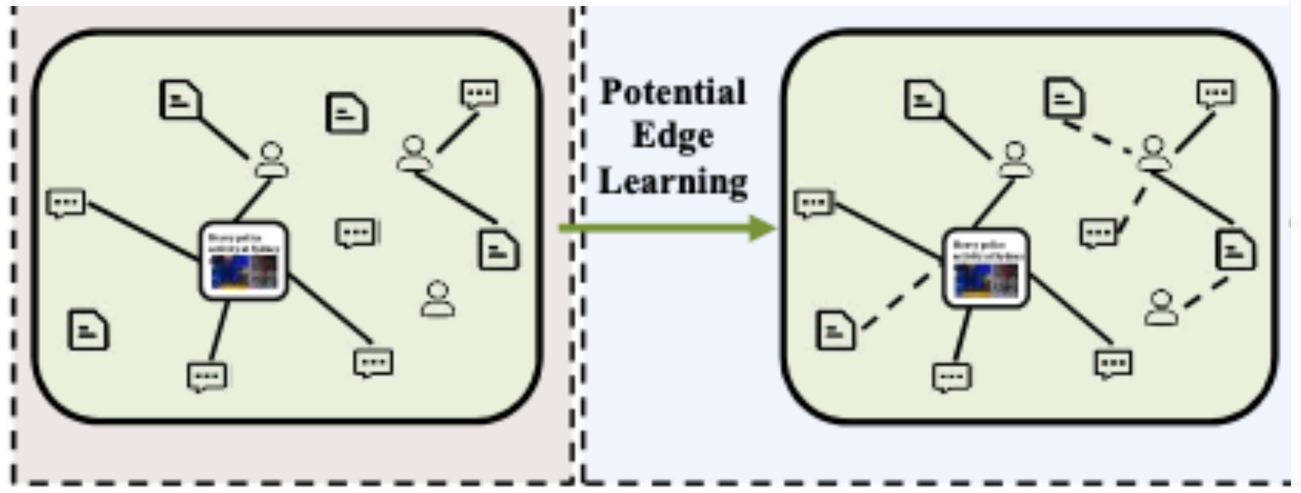


- Text-CNN
- ResNet50

Methodology

Inferring Hidden Links

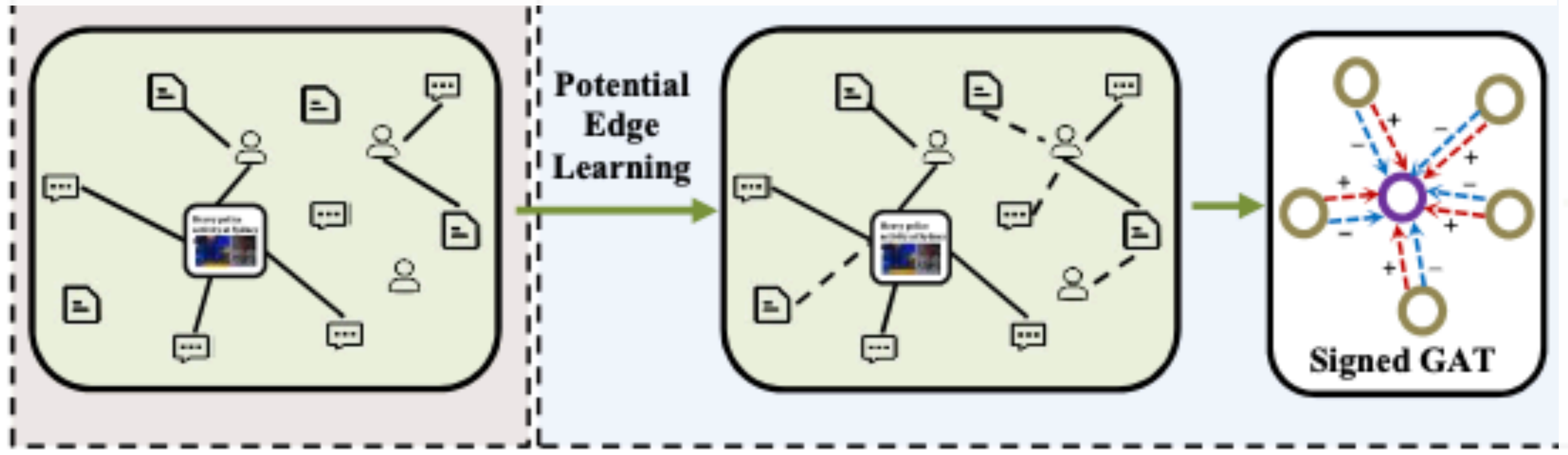
- Update the adjacency matrix $A \in \mathbb{R}^{|V| \times |V|}$
- Correlation between two nodes: $\beta_{ij} = \dfrac{x_i \cdot x_j}{\|x_i\| \, \|x_j\|}$
- $e_{ij} = \begin{cases} 0, & \text{if } \beta_{ij} < 0.5 \\ 1, & \text{otherwise} \end{cases}$
- $a'_{ij} = \begin{cases} 0, & \text{if } e_{ij} = 0 \text{ and } a_{ij} = 0 \\ 1, & \text{otherwise} \end{cases}$



Methodology

Capturing Multi-aspect Neighborhood Relations

- Aim to capture the social graph structure information with GAT.
- Different from conventional GAT, introduce the signed attention mechanism.
  - To capture the positive and negative correlations between neighboring nodes to obtain better graph features.
  - Start from vanilla GAT and then propose signed attention based GAT to capture the multi-aspect correlations.



Methodology

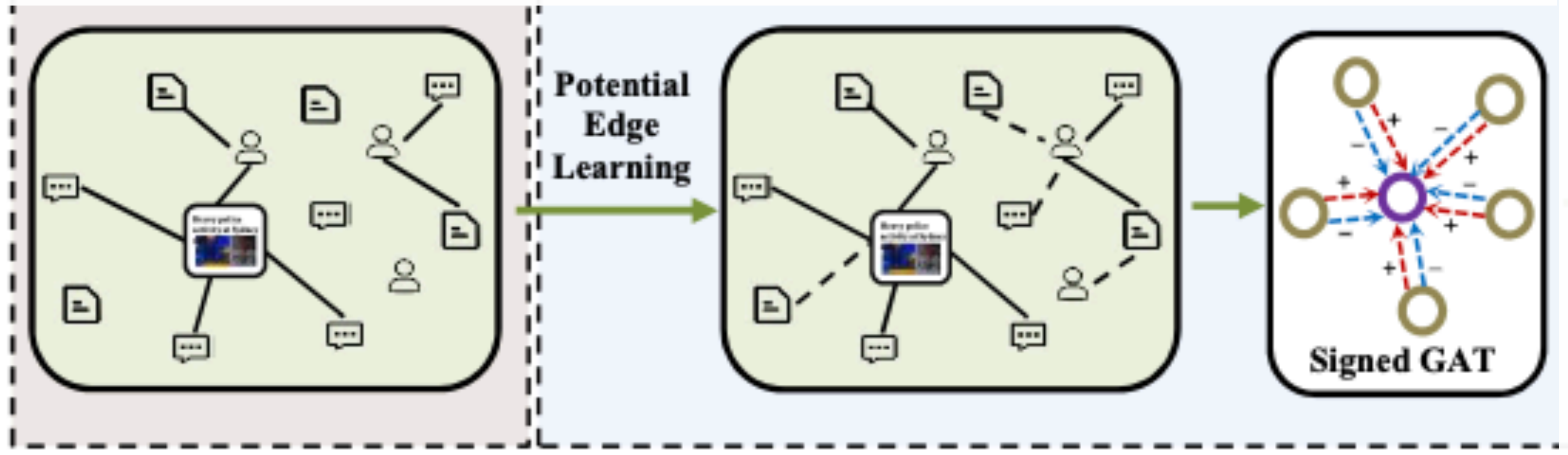
Capturing Multi-aspect Neighborhood Relations

Node n_i

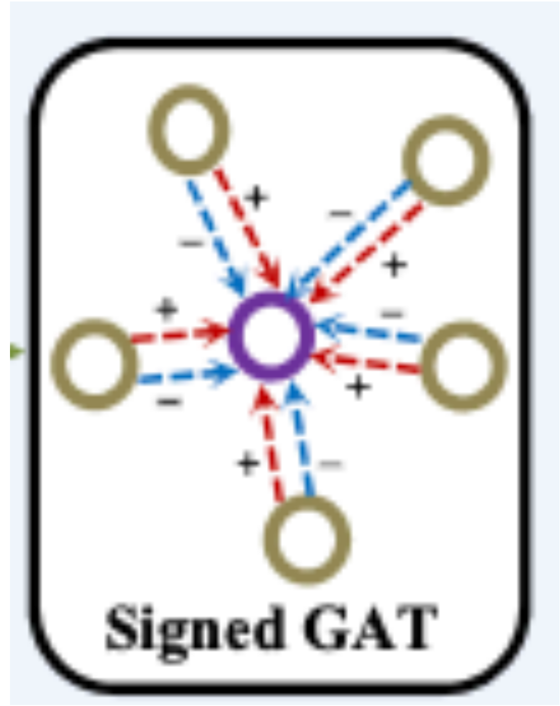
Its neighbor node set $\mathcal{N}_i = \{n'_1, n'_2, \cdots, n'_{|\mathcal{N}_i|}\}$

Attention weights $\mathcal{E}_i = \{e'_{i1}, e'_{i2}, \cdots, e'_{i|\mathcal{N}_i|}\}$

- First calculate the set of attention weights $\mathcal{E}_i$ between n_i and each node in $\mathcal{N}_i$ by
  - $e'_{ij} = \mathrm{LeakyReLU}(\hat{a}[Wx_i || Wx'_j])$
- Existing GATs ignore the negative correlations, to address this issue, design the signed attention based GAT, namely Signed GAT.



Methodology
Signed GAT



Node n_i

Its neighbor node set $\mathcal{N}_i = \{n'_1, n'_2, \cdots, n'_{|\mathcal{N}_i|}\}$

Attention weights $\mathcal{E}_i = \{e'_{i1}, e'_{i2}, \cdots, e'_{i|\mathcal{N}_i|}\}$

- It uses signed attention to involve both the positive and negative relationships between nodes.

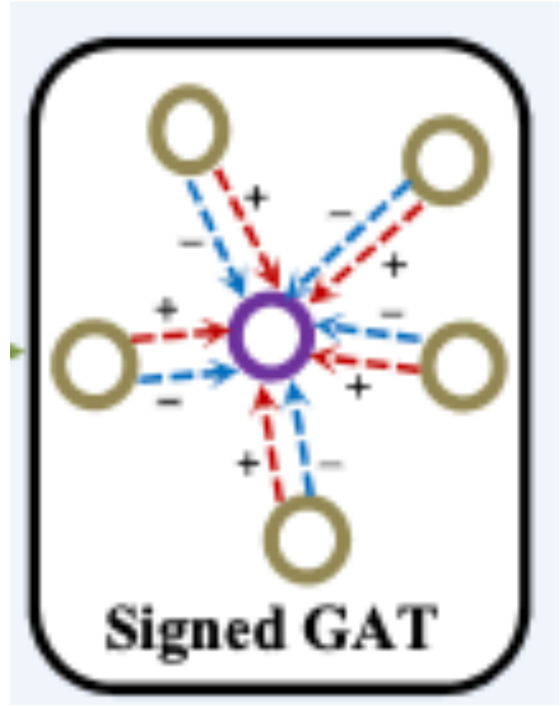
- Specifically, for n_i, denote the inversion of the attention weights $\mathcal{E}_i$ as $\tilde{\mathcal{E}}'_i = -\mathcal{E}_i$.

- Then calculate the normalized weights for both $\mathcal{E}_i$ & $\tilde{\mathcal{E}}_i$ with the softmax function.

- $$\mathcal{E}'_i = \mathrm{softmax}(\mathcal{E}_i)$$
$$\tilde{\mathcal{E}}'_i = \mathrm{softmax}(\tilde{\mathcal{E}}_i)$$

Methodology

Signed GAT



Node n_i

Its neighbor node set $\mathcal{N}_i = \{n'_1, n'_2, \cdots, n'_{|\mathcal{N}_i|}\}$

Attention weights $\mathcal{E}_i = \{e'_{i1}, e'_{i2}, \cdots, e'_{i|\mathcal{N}_i|}\}$

- In order to capture both positive and negative relations between nodes.
  - Utilize the $\mathcal{E}'_i$ & $-\tilde{\mathcal{E}}'_i$ to obtain the weighted sum of the neighbor nodes' features.
  - Then concatenate the two vectors together and pass it through a full connected layer to obtain the final node feature.
- For instance, the node feature of n_i can be obtained by
- $$\hat{x}_i = \sigma(W_n * (\mathcal{E}'_i * X_j || - \tilde{\mathcal{E}}'_i * X_j))$$

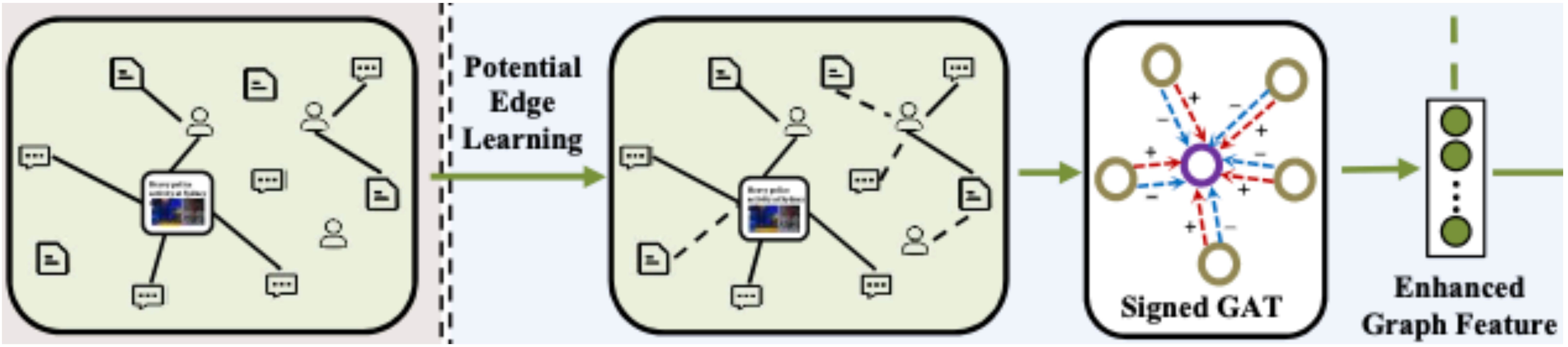
Active function

Weight matrix of the fully connected layer

feature matrix of $\mathcal{N}_i$

Methodology

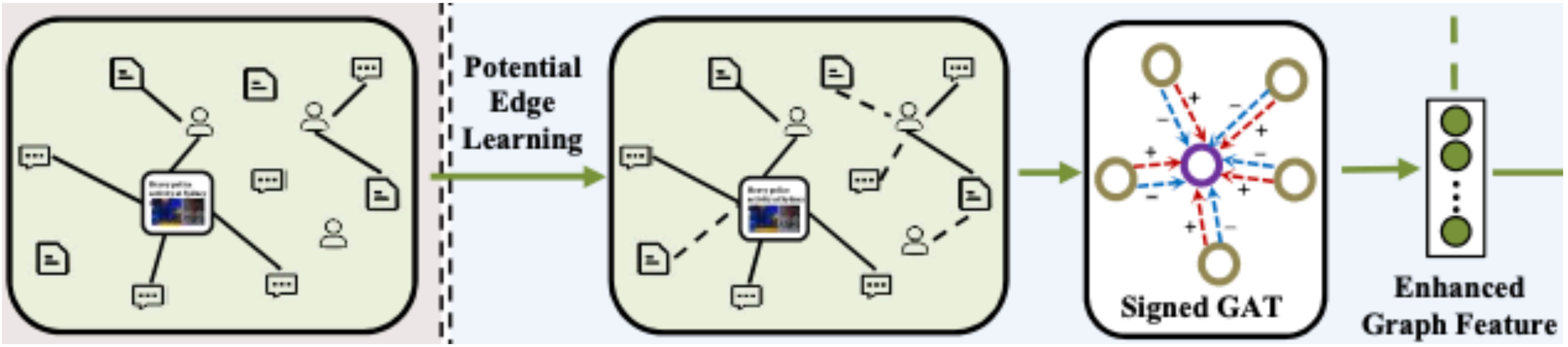
Graph Feature Extractor



- Firstly, enhance the original social graph by augmenting the inferred potential edges, and initialize three types of nodes in the graph.
  - Post, comment nodes
    - Use their sexual features as the initial embeddings.
  - User nodes
    - Use the average of their post and comment embeddings as the initial embeddings to reflect the user characteristics.

Methodology

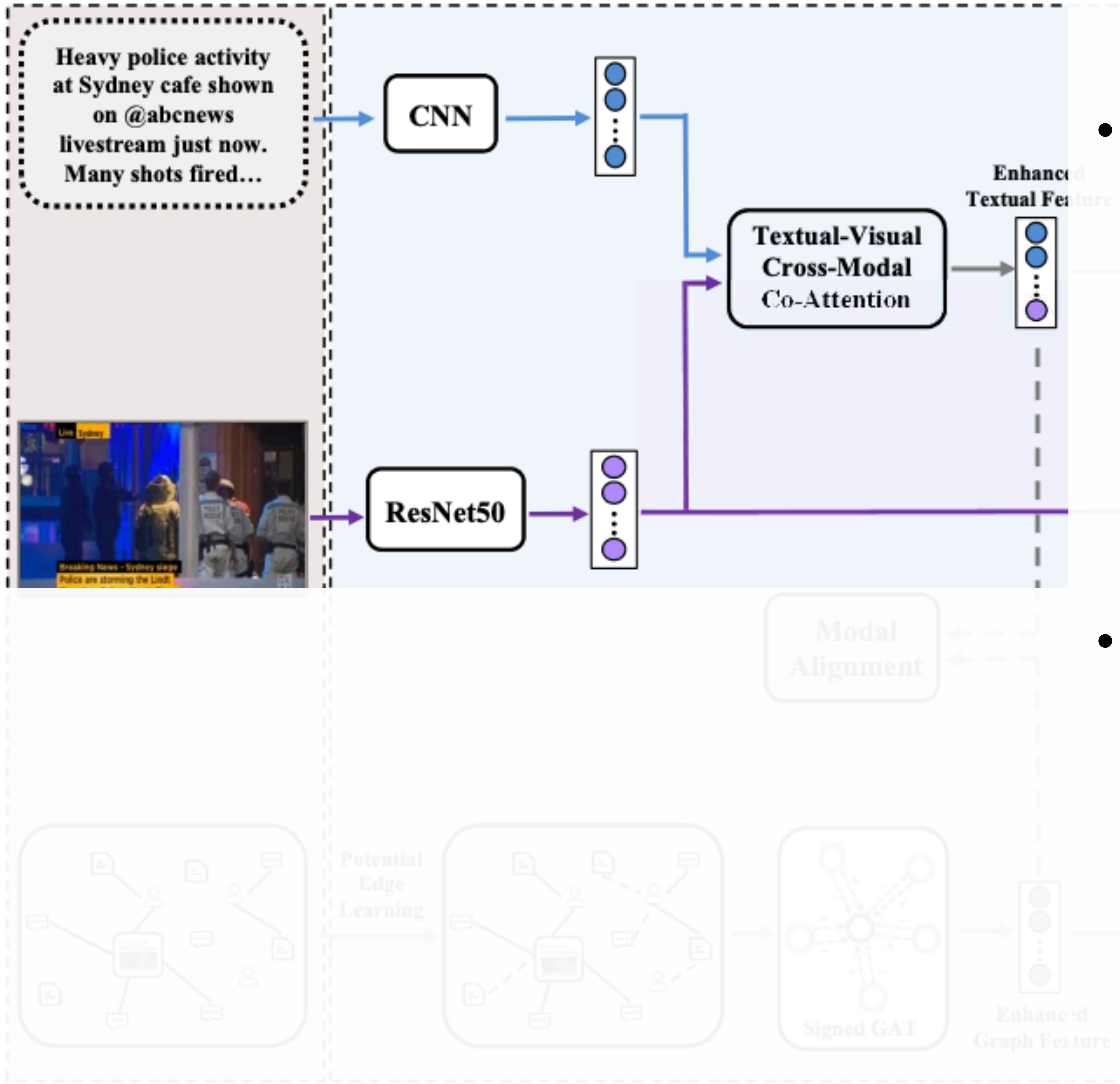
Graph Feature Extractor



- Then use Signed GAT to extract graph structure features from the enhanced social graph.
- For each node, update its embedding and obtain the updated node embedding matrix.
- Then a multi-head attention mechanism is adopted to capture features from different perspectives.
- Concatenate the updated node embeddings of each head together as overall graph feature:
  - $\hat{G} = ||_{h=1}^{H} \sigma(\hat{X}_h)$

Methodology

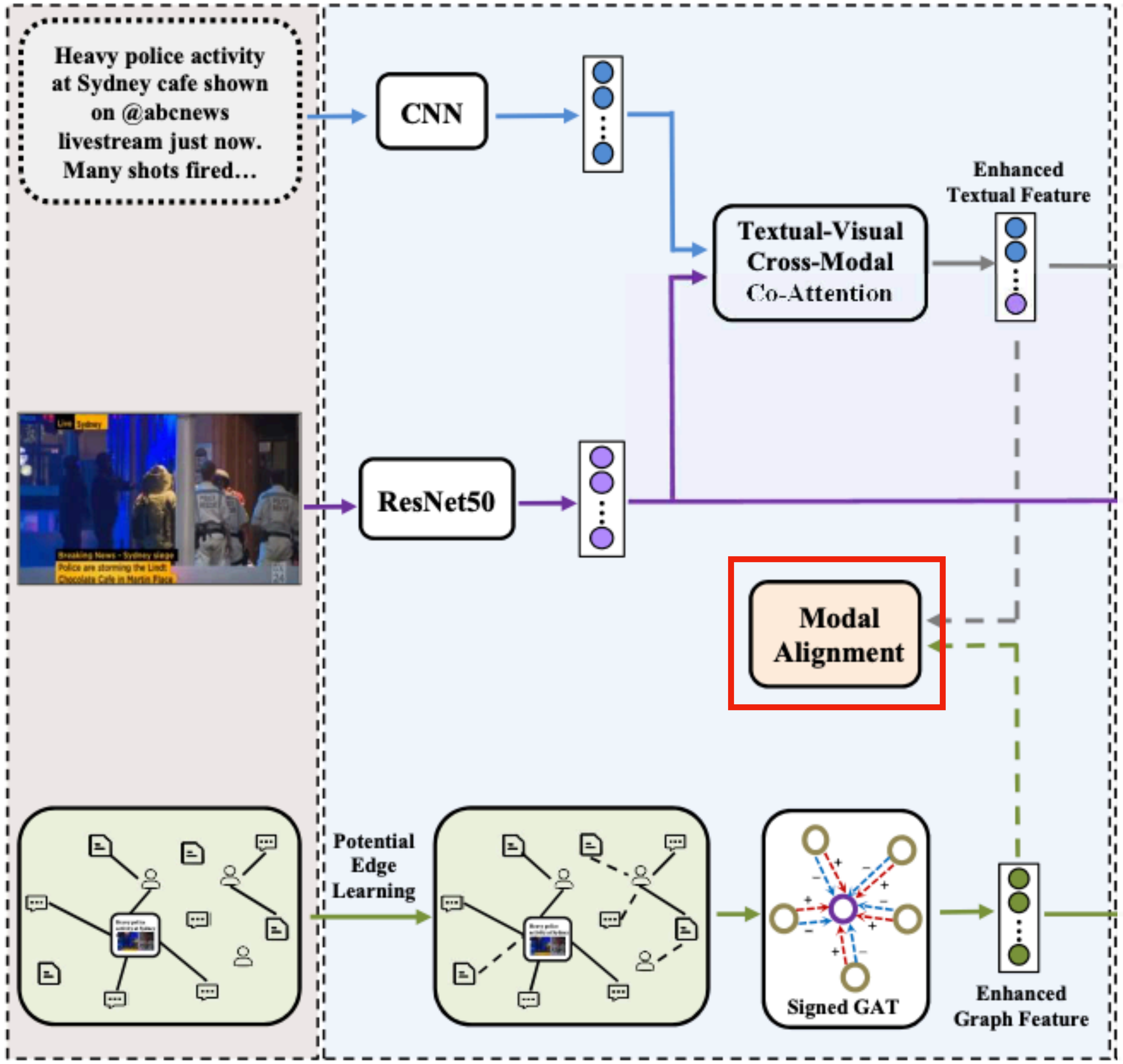
Cross-modal Co-attention Mechanism



- Intra-modal feature representation
  - $$Z_t^i = (\|_{h=1}^{H} \operatorname{softmax}(\frac{Q_t^i K_t^i}{\sqrt{d}}) V_t^i) W_t^O$$
  - $Q_t^i = R_t^i W_t^Q, K_t^i = R_t^i W_t^K, V_t^i = R_t^i W_t^V$
- Cross-modal enhanced feature
  - $$Z_{vt}^i = (\|_{h=1}^{H} \operatorname{softmax}(\frac{Q_v^i K_t^i}{\sqrt{d}}) V_t^i) W_{vt}^O$$
  - $Q_v^i = Z_v^i W_v^Q, K_t^i = Z_t^i W_t^K, V_t^i = Z_t^i W_t^V$

Methodology

Multi-modal Alignment



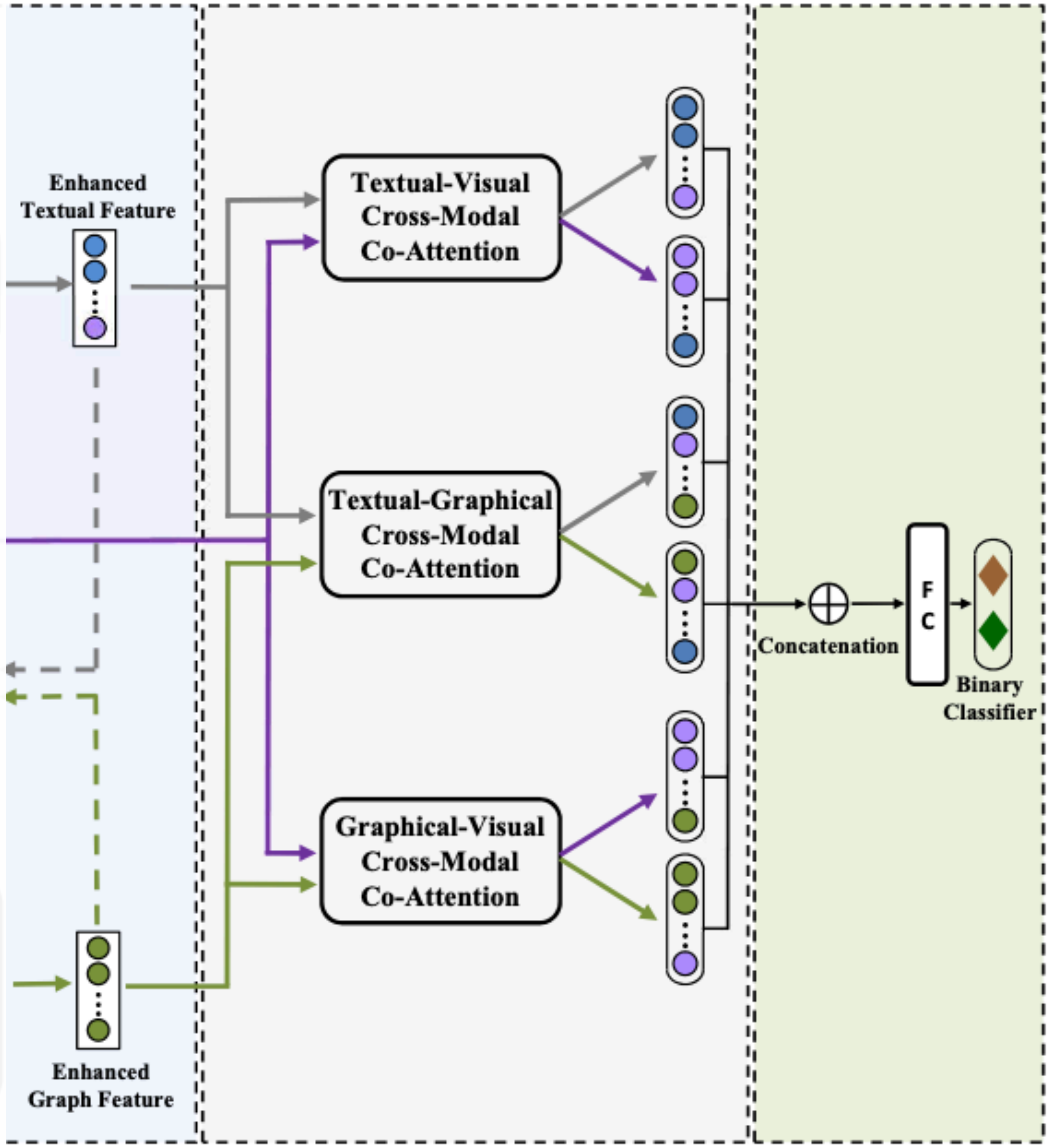
- Enforcing the enhanced textual feature of the post close to its enhanced graphical features in order to refine the representations learned in each modality.
- $$Z_g^{i'} = W_g' Z_g^i$$ $$Z_t^{i'} = W_t' Z_{vt}^i$$
- Then narrow the distance between $Z_g^{i'}$ and $Z_t^{i'}$ with the MSE loss for modal alignment:
- $$\mathcal{L}_{align} = \frac{1}{n} \sum_{i=1}^{n} (Z_g^{i'} - Z_t^{i'})^2$$

Methodology

Fusing the Above Multi-modal Features

- Again perform the aforementioned cross-modal co-attention mechanism among each pair of the three modal features.

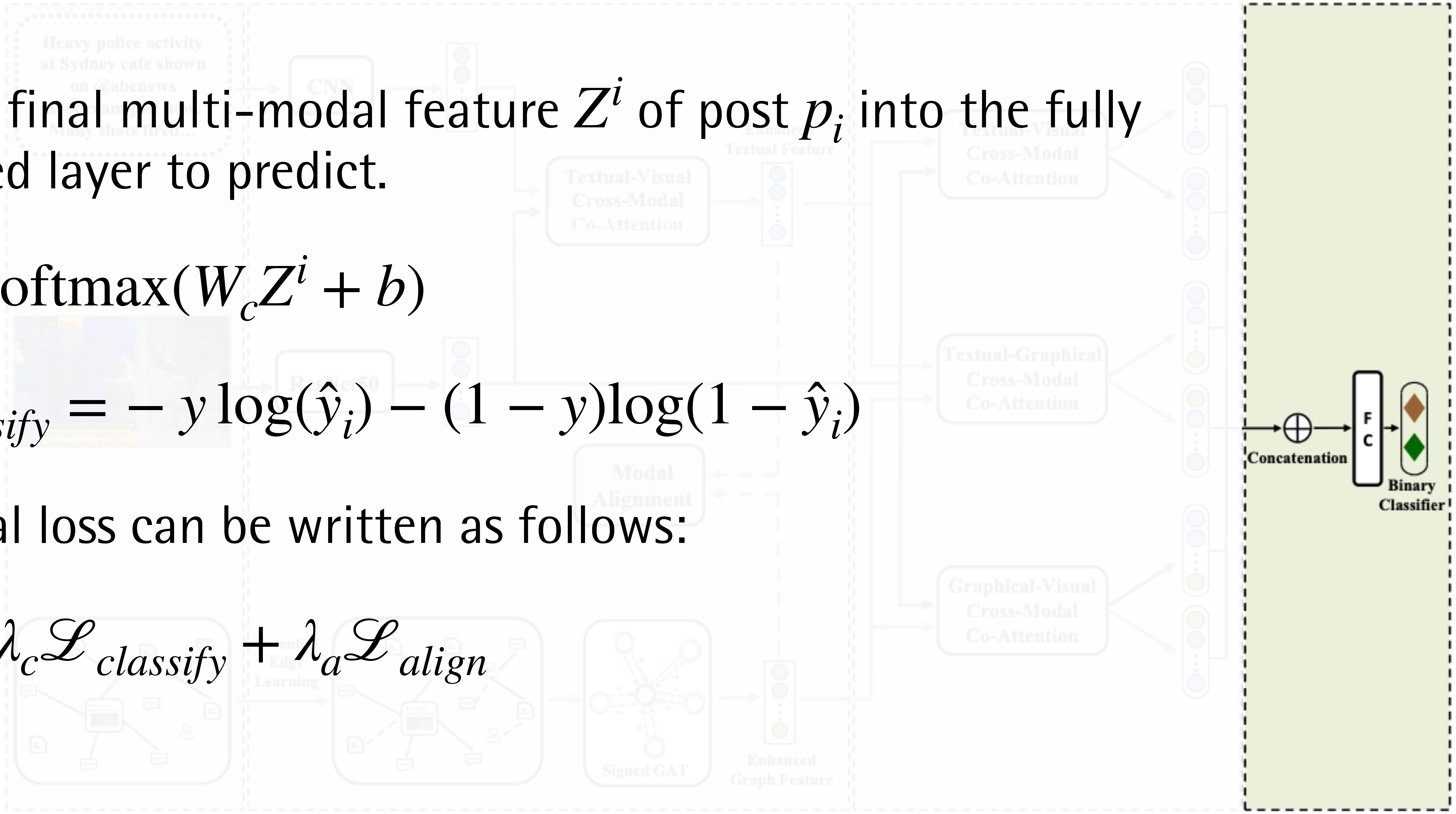
- Finally get six cross-modal enhanced features then concatenate them:
  - $Z^i = \mathrm{concat}(\tilde{Z}^i_{tv}, \tilde{Z}^i_{vt}, \tilde{Z}^i_{gt}, \tilde{Z}^i_{tg}, \tilde{Z}^i_{gv}, \tilde{Z}^i_{vg})$



Methodology

Classification with Adversarial Training

- Feed the final multi-modal feature Z^i of post p_i into the fully connected layer to predict.
  - $\hat{y}_i = \mathrm{softmax}(W_c Z^i + b)$
  - $\mathscr{L}_{classify} = -y\log(\hat{y}_i) - (1-y)\log(1-\hat{y}_i)$
- Then final loss can be written as follows:
  - $\mathscr{L} = \lambda_c \mathscr{L}_{classify} + \lambda_a \mathscr{L}_{align}$



Experiments

Datasets

Statistic	Non-rumors	False Rumors	Images	Users	Comments
PHEME	1428	590	2018	894	7388
Weibo	877	590	1467	985	4534

- Weibo
- PHEME
- Train : Valid : Test = 7:1:2

- $\mathscr{L} = \lambda_c \mathscr{L}_{classify} + \lambda_a \mathscr{L}_{align}$
- $\lambda_c = 2.15,\ \lambda_a = 1.55$

Experiments

Baselines

- Textual & Visual features: EANN, MVAE, SAFE
- Textual only
  - QSAN: integrates the quantum-driven text encoding and a novel signed attention mechanism for false information detection.
- Social graphical features
  - EBGCN: rethinks the reliability of latent relations in the propagation structure by adopting a Bayesian approach.
  - GLAN: jointly encodes the local semantic and global structural information and applies a global-local attention network for rumor detection.

Experiments

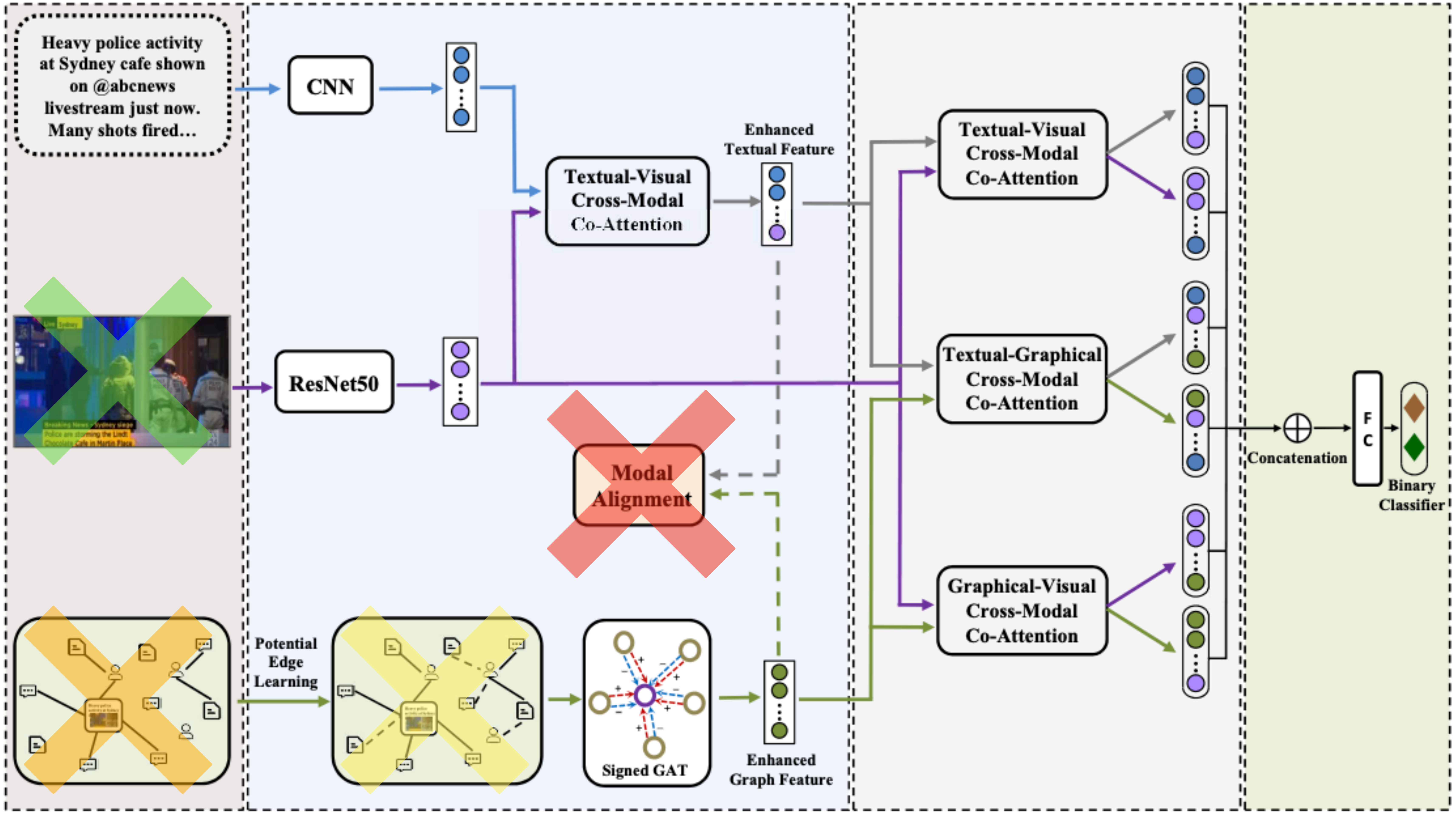
Result & Analysis

	PHEME				Weibo			
Method	Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
EANN	77.13±0.96	71.39±1.07	70.07±2.19	70.44±1.69	80.96±2.26	80.19±2.37	79.68±2.46	79.87±2.40
MVAE	77.62±0.64	73.49±0.81	72.25±0.90	72.77±0.81	71.67±0.89	70.52±0.95	70.21±1.01	70.34±0.98
QSAN	75.13±1.19	69.97±2.03	65.80±1.72	66.87±1.70	71.01±1.81	71.02±0.95	67.54±3.27	67.58±3.59
SAFE	81.49±0.84	79.88±1.22	79.50±0.81	79.68±0.70	84.95±0.85	84.98±0.82	84.95±0.91	84.96±0.86
EBGCN	82.99±0.65	81.31±0.73	79.29±0.71	79.82±0.64	83.14±2.01	85.46±2.12	81.76±1.54	81.45±1.74
GLAN	83.32±1.64	81.25±2.06	77.13±3.26	78.51±2.68	82.44±2.02	82.45±2.26	80.86±1.71	81.26±1.93
MFAN	**88.73±0.83**	**87.07±1.41**	**85.61±1.65**	**86.16±1.04**	**88.95±1.43**	**88.91±1.60**	**88.13±1.68**	**88.33±1.53**

- For the methods that consider both textual and visual information.
  - SAFE outperforms other methods, indicating the importance of considering interactions between modalities.
- GLAN & EBGCN outperform most other methods.
  - Indicating that the social graph information is beneficial for rumor detection.
- MFAN significantly outperforms all the other approaches.
  - Demonstrating that considering visual, latent links, and modal alignment can further improve the performance.

Experiments
Ablation Analysis

Method		-w/o V	-w/o G	-w/o P	-w/o A	MFAN
PHEME	Acc.	85.66	86.29	86.91	87.12	**88.73**
	F1.	82.47	82.15	83.93	84.41	**86.16**
Weibo	Acc.	84.14	85.08	86.17	86.98	**88.95**
	F1.	83.88	84.48	85.44	86.42	**88.33**



Conclusion
of MFAN

- Propose a multi-modal rumor detection framework.
- Incorporates three types of modalities. (text, image, and social graph)
- To improve the social graph feature learning, both the graph topology and neighborhood aggregation procedure are enhanced based on GAT.
- Proposed framework enables more effective multi-modal fusion by introducing cross-modal alignment.

Comments

of MFAN

- Need network data to enhance performance.
- Utilized co-attention module.