Hierarchical Multi-modal Contextual Attention Network for Fake News Detection

Shengsheng Qian
National Lab of Pattern Recognition, Institute of Automation, CAS
University of Chinese Academy of Sciences
shengsheng.qian@nlpr.ia.ac.cn

Jinguang Wang
HeFei University of Technology
wangjinguang502@gmail.com

Jun Hu
National Lab of Pattern Recognition, Institute of Automation, CAS
hujunxianligong@gmail.com

Quan Fang
National Lab of Pattern Recognition, Institute of Automation, CAS
University of Chinese Academy of Sciences
qfang@nlpr.ia.ac.cn

Changsheng Xu
National Lab of Pattern Recognition, Institute of Automation, CAS
University of Chinese Academy of Sciences
Peng Cheng Laboratory
csxu@nlpr.ia.ac.cn

SIGIR'21

220725 Chia-Chun Ho

Outline
of HMCAN

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction

Fake News Detection

- Social media websites are convenient platforms for people to share information, express and exchange opinions in their daily life.
- However, the authenticity of these information is difficult to guarantee since users do not check the reliability of the shared information.
  - Which has led to the widespread dissemination of considerable fake news.
- Therefore, detecting fake news on social media websites to ensure that users obtain true information has become a top priority.

Introduction

Existing Approaches

- Traditional learning methods
  - Design plenty of hand-crafted features from the media content of posts and the social content of users.
  - SVM, Decision Tree… etc.
  - However, the content of fake news is highly complicated and hard to be fully captured by hand-crafted features.

Introduction
Existing Approaches

- Deep learning methods
  - Many multi-modal representation methods utilize deep schemes to learn the representative features, and obtain superior performance for fake news detection.
  - MVAE, SAME, SpotFake, SpotFake+… etc.
  - Although these approaches show promising performance on fake news detection tasks.
    - They are still insufficient to take advantage of the multi-modal context information and the hierarchical semantics of text content.

Introduction
Challenges (1/2)

- How to fully utilize the multi-modal context information and extract high-order complementary information from it to enhance the performance of fake news detection?
  - The visual content of news posts usually contain may uncertain elements that are difficult to understand without the help of the text information.
  - The components they employ to capture multi-modal context are too simple to extract high-order complementary information from the multi-modal context.

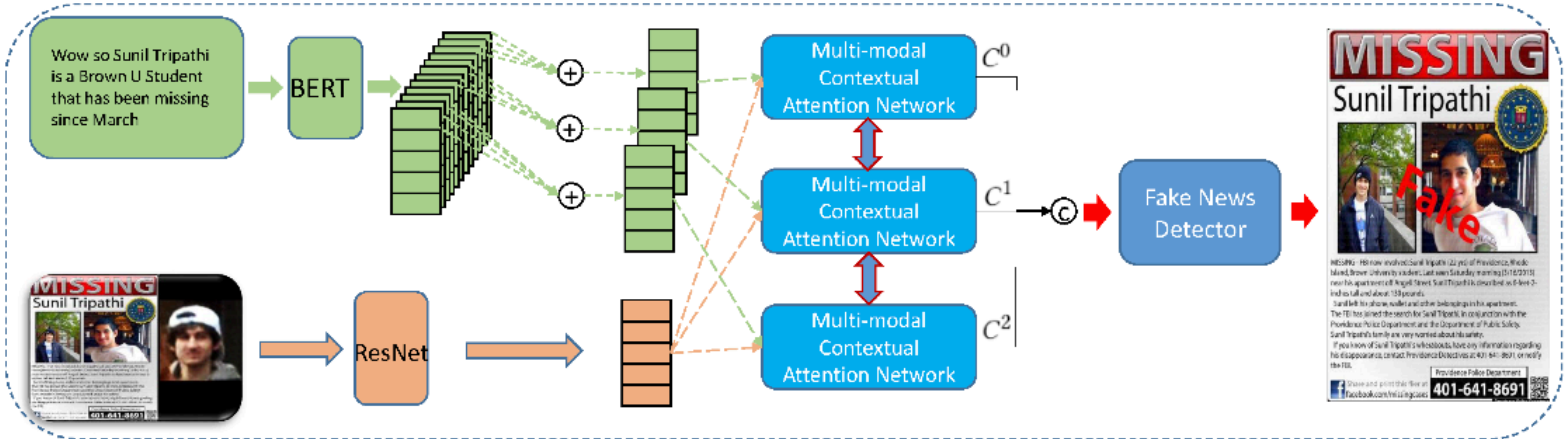
Introduction

Challenges (2/2)

- How to explore and capture the hierarchical semantics of text information to learn a better representation of multi-modal news?
  - Most SOTA models user BERT as text encoders, which can provide hierarchical semantics of text.
    - But most of them only utilize the output of the last layers of these hierarchical models.
    - Some works explore the potential of exploiting the semantic knowledge in the intermediate layers of BERT models, showing that many downstream tasks can benefit from the full hierarchical semantics.

Introduction

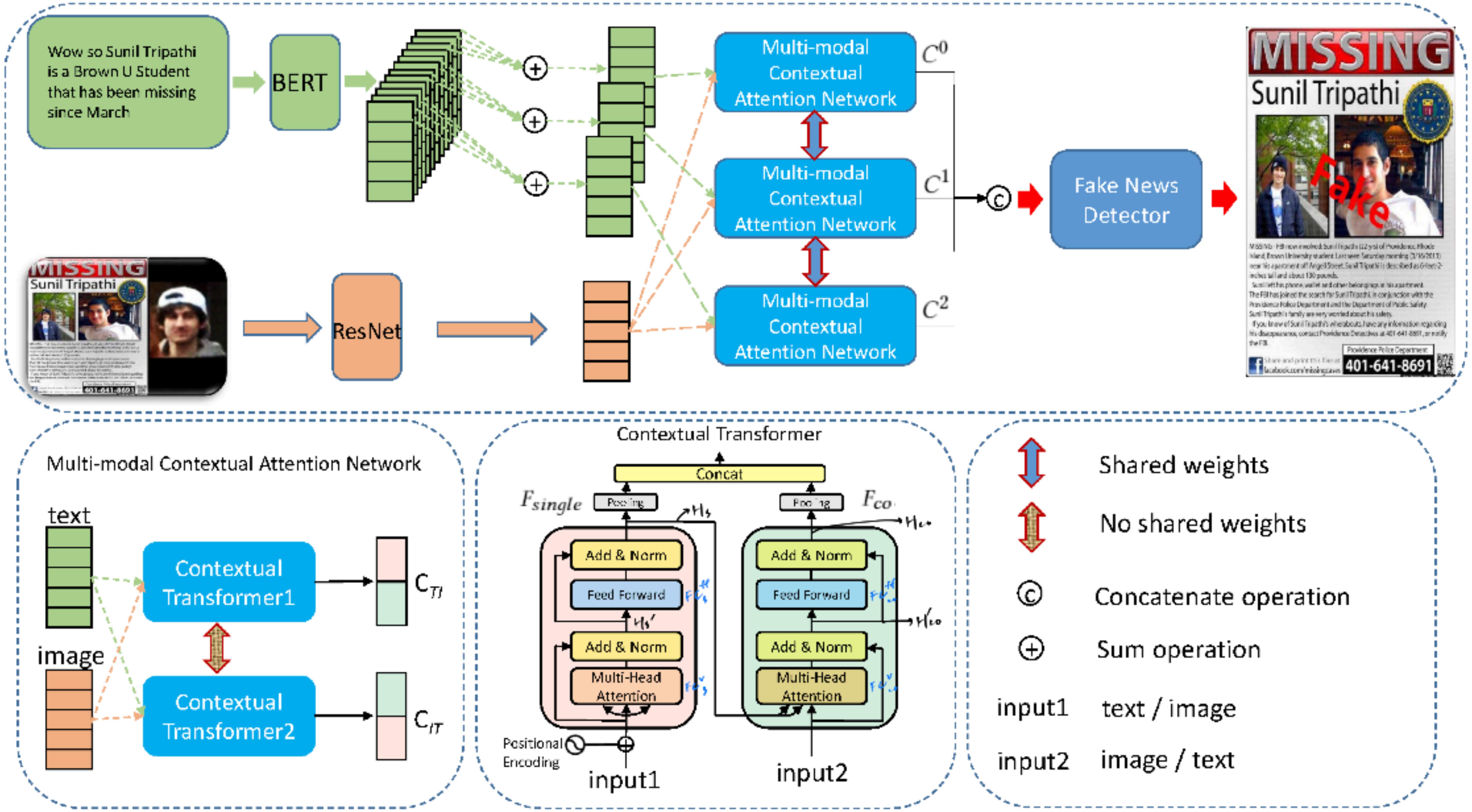
HMCAN & Contributions



- Propose a novel Hierarchical Multi-modal Contextual Attention Network for FND.
  - By jointly modeling the multi-modal context information and the hierarchical semantics of text in a unified deep model.
  - Propose a multi-modal contextual attention network.
    - Modeling the multi-modal context for each news posts, where data from different modalities can complement each other to provide a better understanding of the multi-modal data.
  - Design a hierarchical encoding network.
    - For capture the rich hierarchical semantics for fake news detection.

Methodology

Hierarchical Multi-modal Contextual Attention Network (HMCAN)



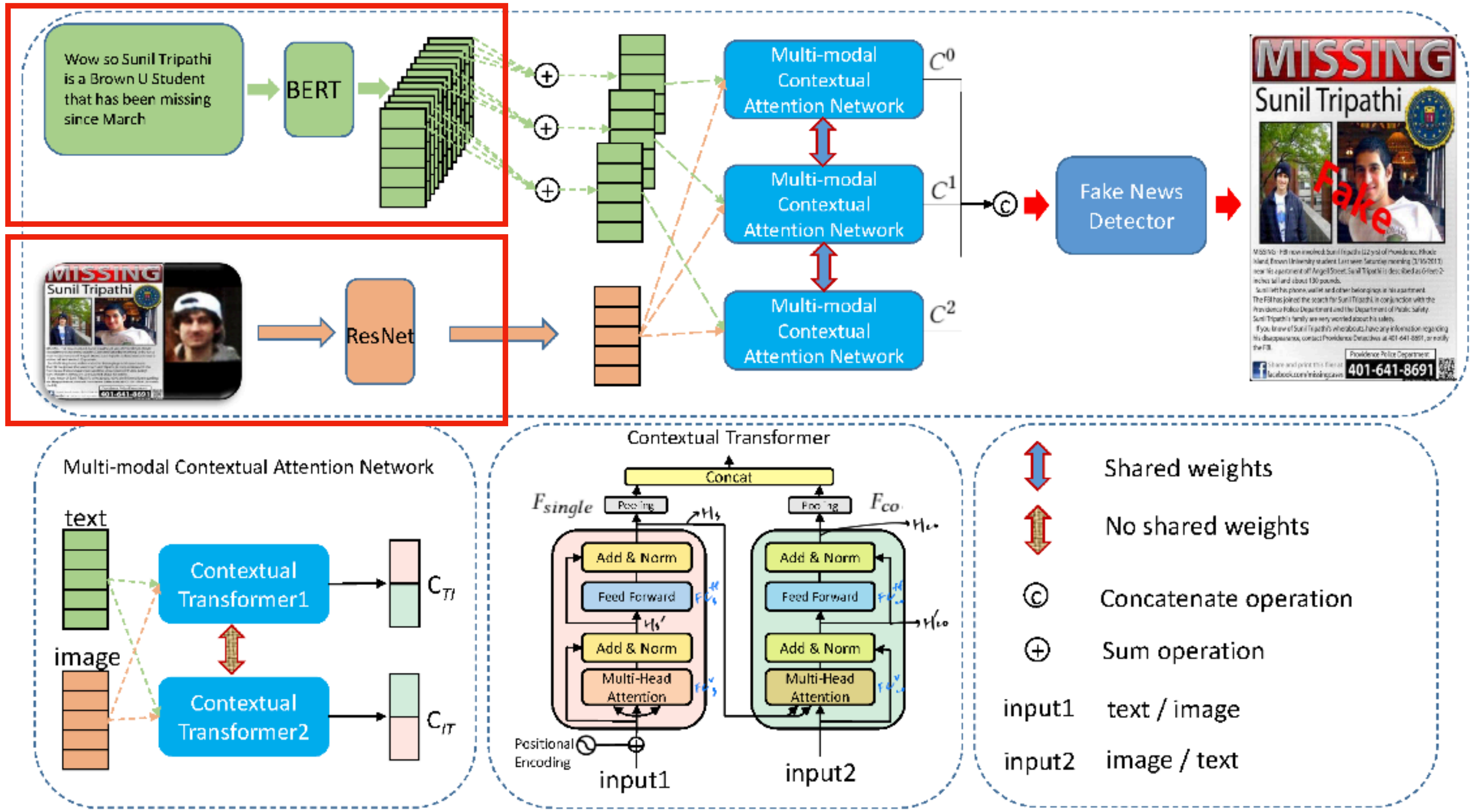
Methodology

Problem Definition

- P: a multi-modal post from social media consisting of text messages and corresponding images.
- The model will output $Y = \{0,1\}$ to indicate to the label of the post.
  - $Y = 0$: real
  - $Y = 1$: fake

Methodology

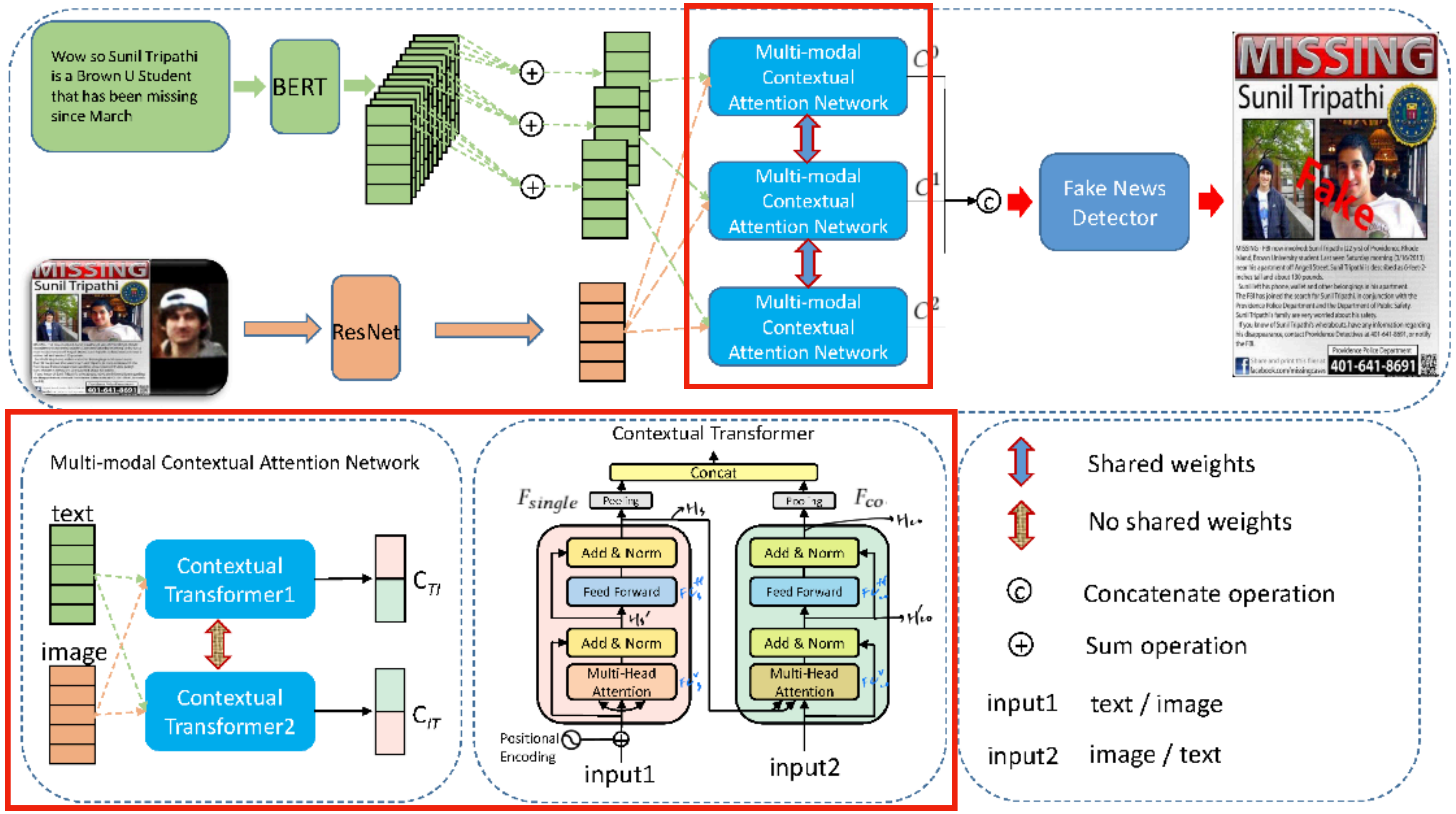
Text and Image Encoding Network



- Text Encoding Network
  - BERT
- Image Encoding Network
  - ResNet-50

Methodology

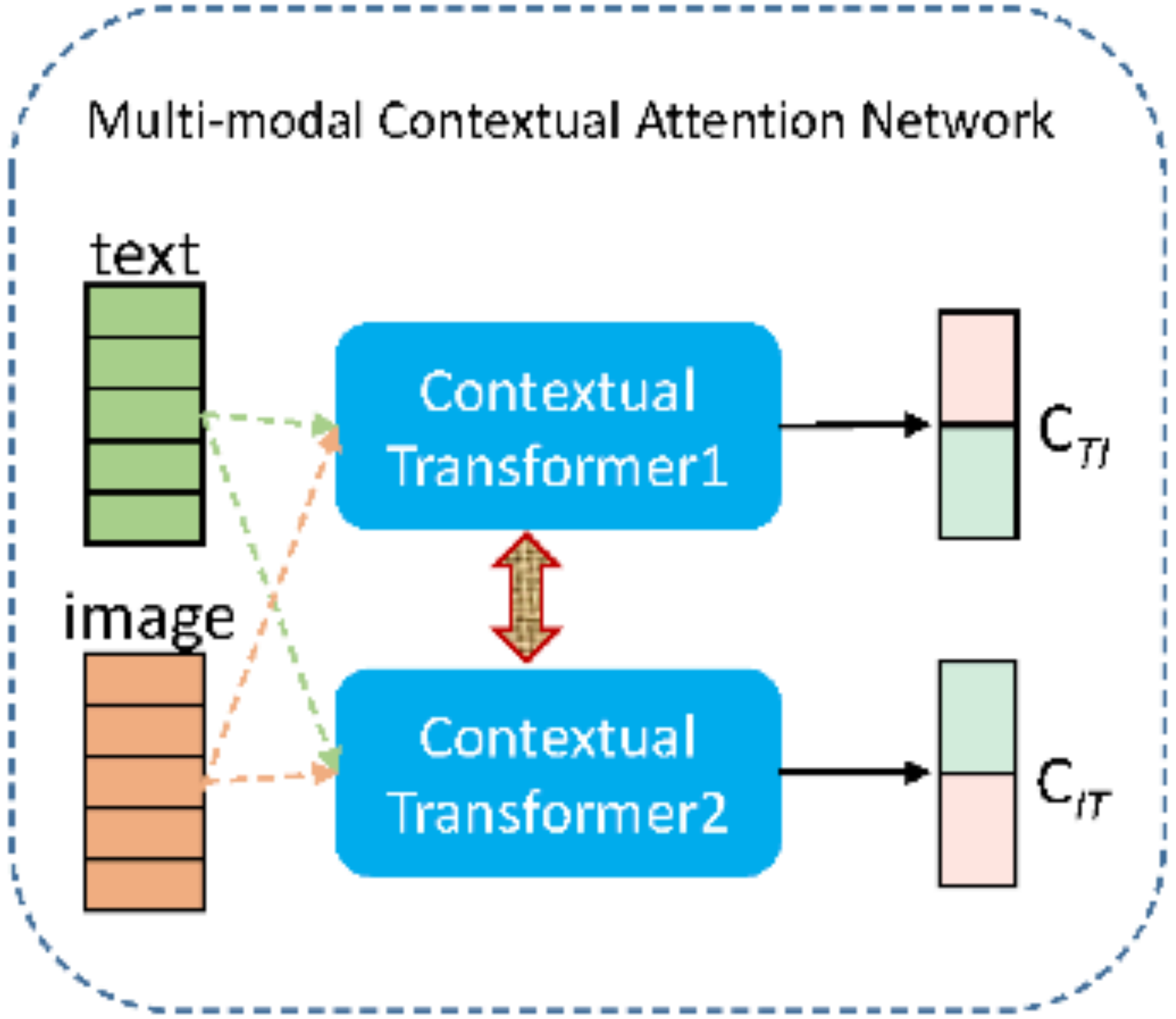
Multi-modal Contextual Attention Network



- Build the multi-modal context information.
- Extract high-order complementary information.

Methodology

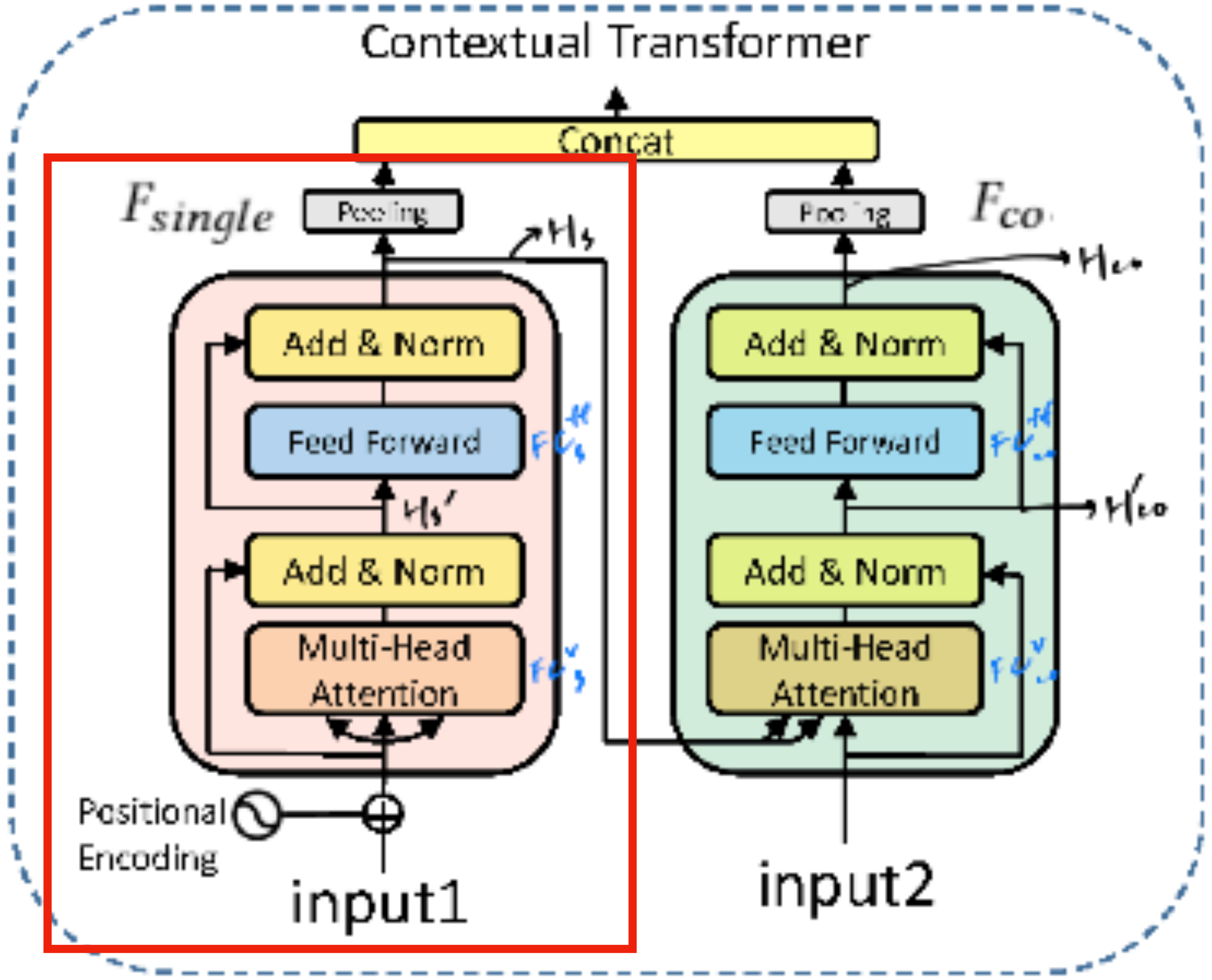
Multi-modal Contextual Attention Network



- Multi-modal contextual attention network consists of two contextual transformer units (Contextual Transformer1, 2).
- Each context transformer unit focuses on different context information for multi-modal representation learning.

Methodology

Multi-modal Contextual Attention Network - F_{single}



- Self-attention network F_{single} (the left part) is utilized to learn the representation of text (input1).

- The self-attention network computes a intra-modality affinity matrix A_s.

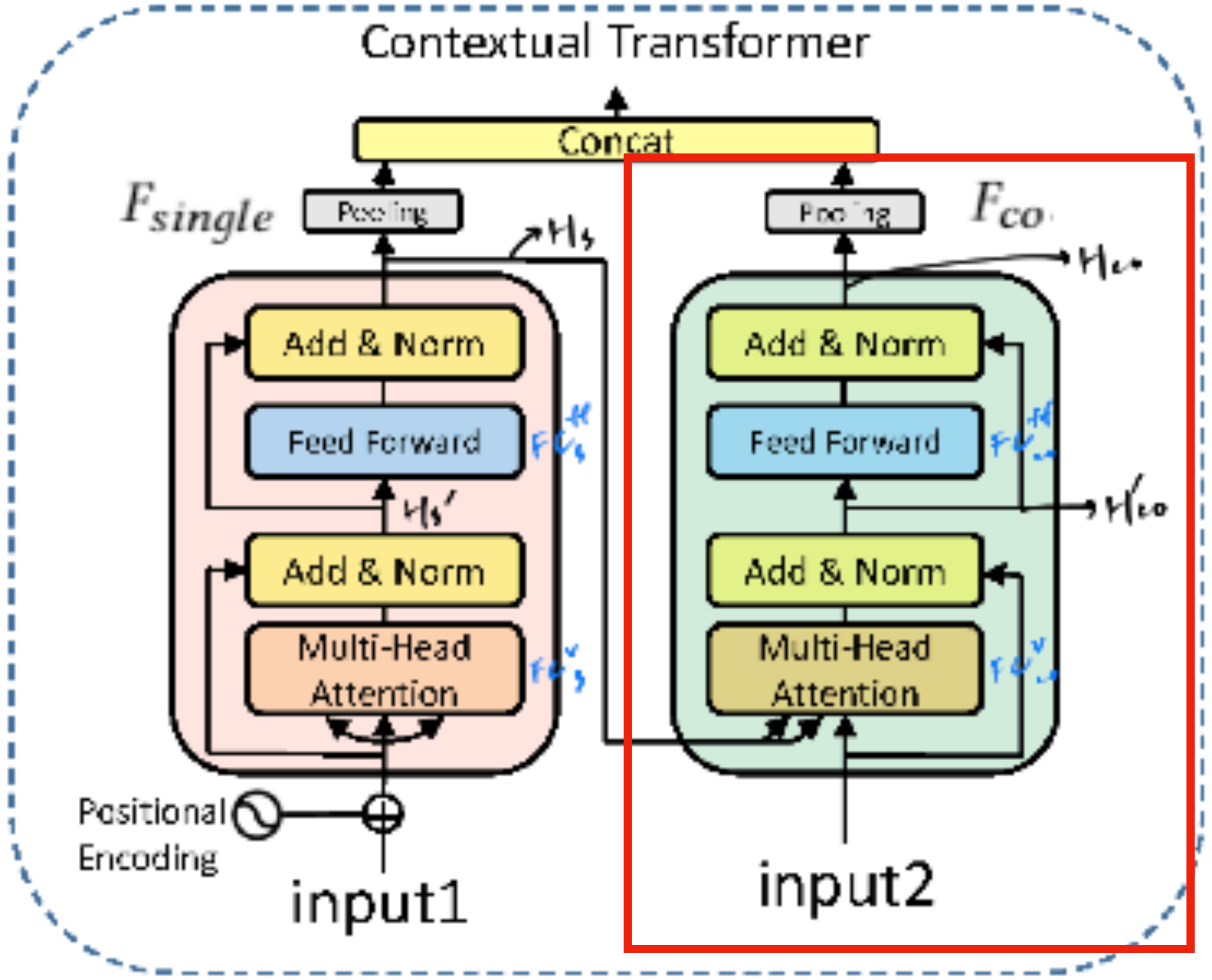
  - $$A_s = \text{softmax}\left(\frac{FC_s^Q(input1) \cdot FC_s^K(input1)^\top}{\sqrt{d}}\right)$$

- Representation of text H_s can be learned as follows:

  - $H_s' = layer\_norm(input1 + A_s \cdot FC_s^V(input1))$
  - $H_s = layer\_norm(H_s' + FC_s^{ff}(H_s'))$

Methodology

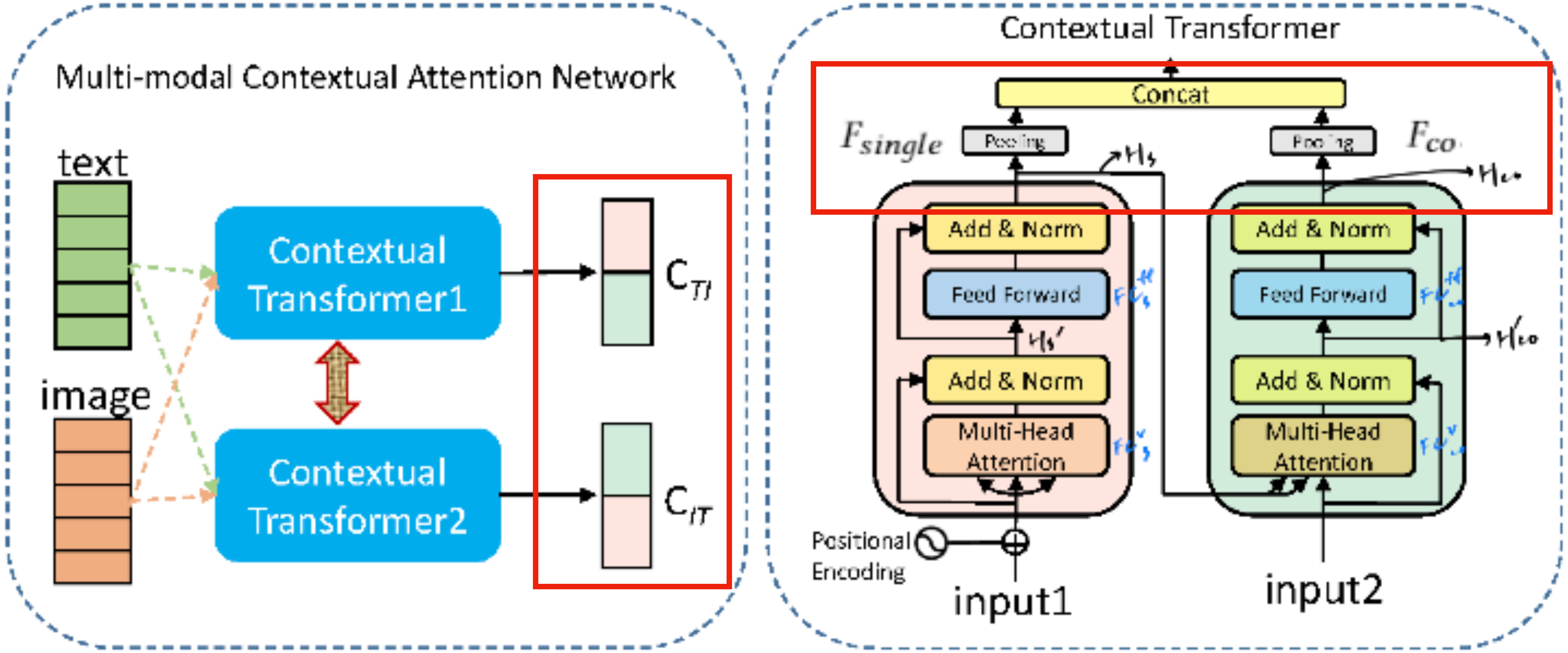
Multi-modal Contextual Attention Network - F_{co}



- The core idea is to extract information that is relevant to the image from the learned text representation, which can complement the visual information.
- F_{co} computes a intermodality affinity matrix A_{co}.
  - $$A_{co} = \mathrm{softmax}\left(\frac{FC^{Q}_{co}(input2) \cdot FC^{K}_{co}(H_s)^{\top}}{\sqrt{d}}\right)$$
- F_{co} learns the multi-modal context-aware text representation H_{co} as follows:
  - $H'_{co} = layer\_norm(input2 + A_{co} \cdot FC^{V}_{co}(H_S))$
  - $H_{co} = layer\_norm(H'_{co} + FC^{ff}_{co}(H'_{co}))$

Methodology

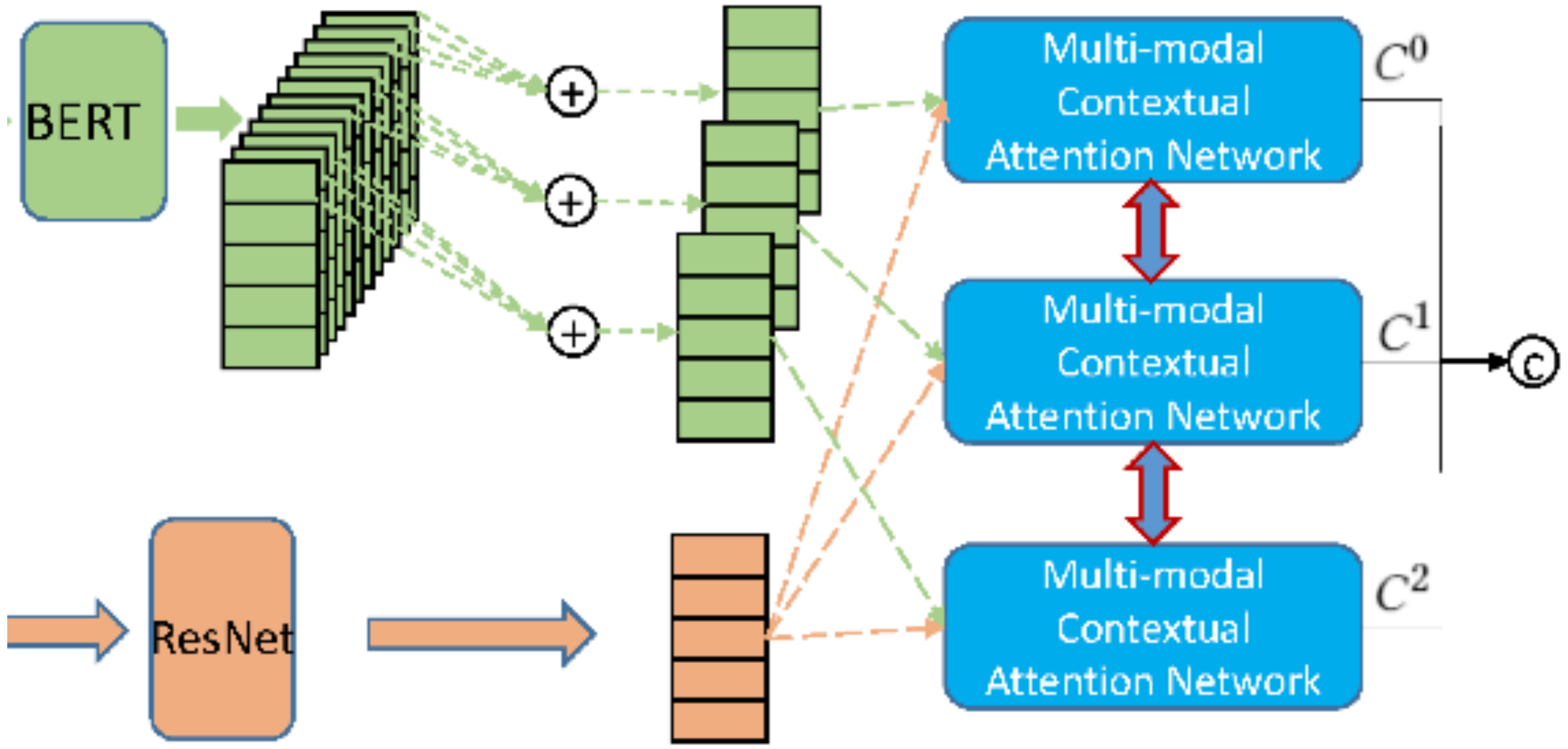
Multi-modal Contextual Attention Network



- H_s and H_{co} are pooled into two feature vectors, which are then concatenated into a feature vector (C_{TI}/C_{IT}) as the multimodal contextual representation of the text.
- Make the output of the multi-modal contextual attention network.
  - $C = \alpha C_{TI} + \beta C_{IT}$
  - $\alpha + \beta = 1$

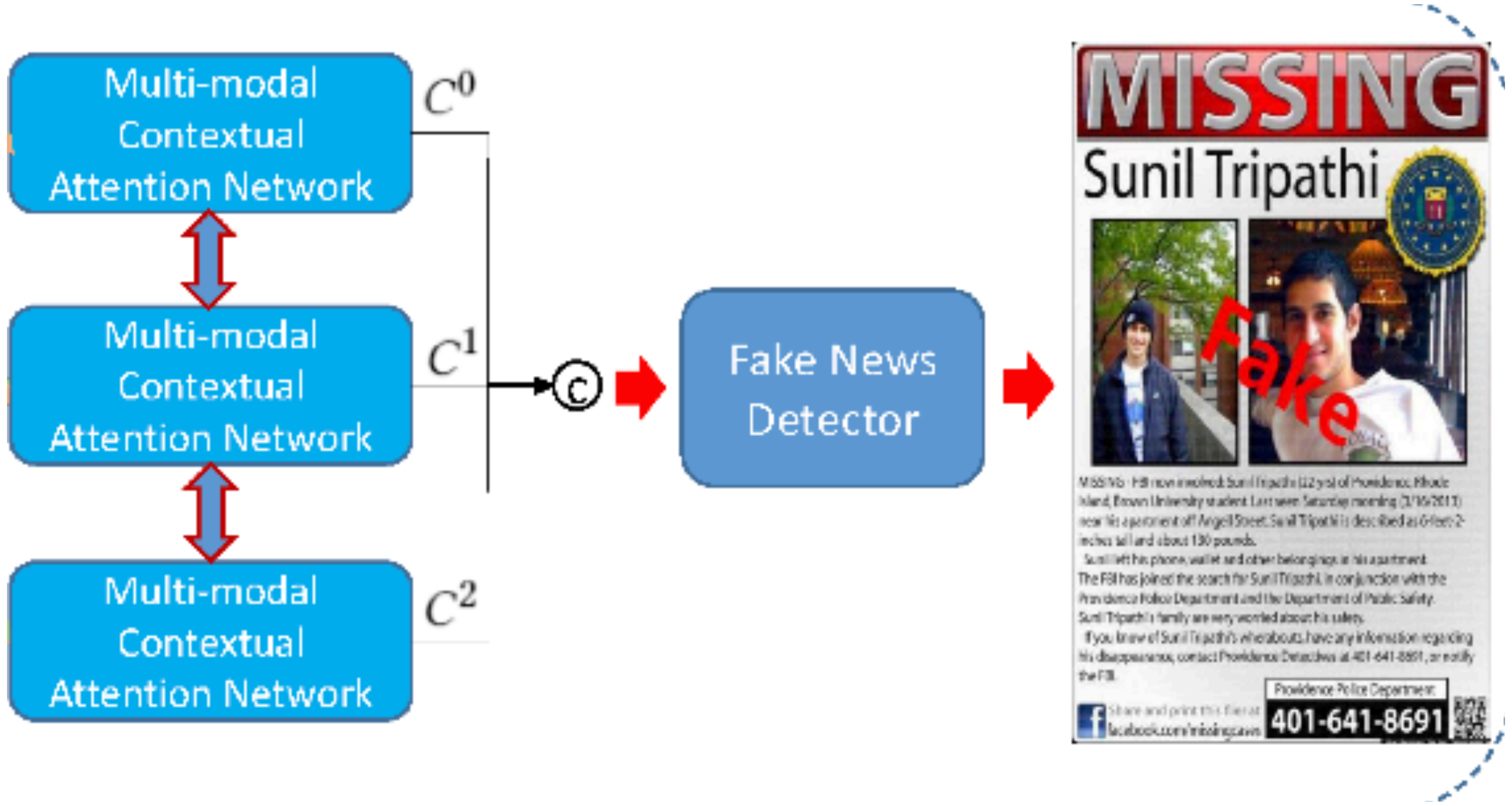
Methodology

Hierarchical Encoding Network



- BERT can provide hierarchical semantics for text, which consists of the outputs of 11 intermediate layers and 1 output layers.
- Reduce the 12 layer outputs into g group outputs by integrating every $12/g$ adjacent layers of BERT
- $\mathbf{s}_\mathbf{i}^0 = \sum_{j=1}^{4} f_B(W)_{j,i}, \mathbf{s}_\mathbf{i}^1 = \sum_{j=5}^{8} f_B(W)_{j,i}, \mathbf{s}_\mathbf{i}^2 = \sum_{j=9}^{12} f_B(W)_{j,i}$
- $C = \text{concat}(C^0, C^1, C^2)$

Methodology
Fake News Detector



- It deploys a fully connected layer with the corresponding activation function.
  - $\hat{P}_n = \sigma(W_f C_n + b)$
- Employ cross entropy loss:
  - $\mathscr{L}(\Theta) = \sum_{n=1}^{N} - [\mathrm{Y}_n \log(\hat{P}_n + (1 - Y_n)\log(1 - \hat{P}_n)]$

Experiments
Datasets

News	WEIBO	TWITTER	PHEME
# of Fake News	4749	7898	1972
# of Real News	4779	6026	3830
# of Images	9528	514	3670

- Weibo
    - Train : Test = 8:2
- Twitter
- PHEME
    - Train : Test = 8:2

Experiments

Baselines (1/2)

- Single-modal
  - SVM-TS, CNN, GRU, TextGCN
- Multi-modal
  - EANN
  - att-RNN
  - MVAE
  - SpotFake, SpotFake+
  - SAFE

Experiments
Result & Analysis

- SVM-TS performs the worst among all.
  - Indicating that the hand-crafted features are weak and insufficient to identify fake news.
- Deep learning models (CNN, GRU) have better performance than SVM-TS.
  - Indicating the superior advantages over traditional methods.
  - In Twitter dataset, CNN fails to capture long-range semantic relationships between words.
  - TextGCN is better than them show that graph structure can effectively capture word co-occurrence and document-word relationships.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	att_RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake*	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
TWITTER	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	att_RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake*	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	SVM-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	att_RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

Experiments

Result & Analysis

- att-RNN has superior performance than GRU.
  - Showing the effectiveness of the attention.
  - It takes into account the text related parts of the image, thus improving the performance of the model.
- MVAE has better performance than single-modal models.
  - Indicates that additional visual information can be used as complementary information.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	att_RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake*	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
TWITTER	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	att_RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake*	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	SVM-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	att_RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

Experiments

Result & Analysis

- SAFE outperforms CNN on the three datasets.
  - Because SAFE jointly uses multi-modal (text and visual) and relational information to learn the representation of posts.
- In addition, SpotFake and SpotFake+ achieve better results on all baselines on Twitter and Weibo datasets.
  - Indicating that the pre-trained BERT and XLNet can obtain better textual information to improve model performance.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	att_RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake*	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
TWITTER	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	att_RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake*	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	SVM-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	att_RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

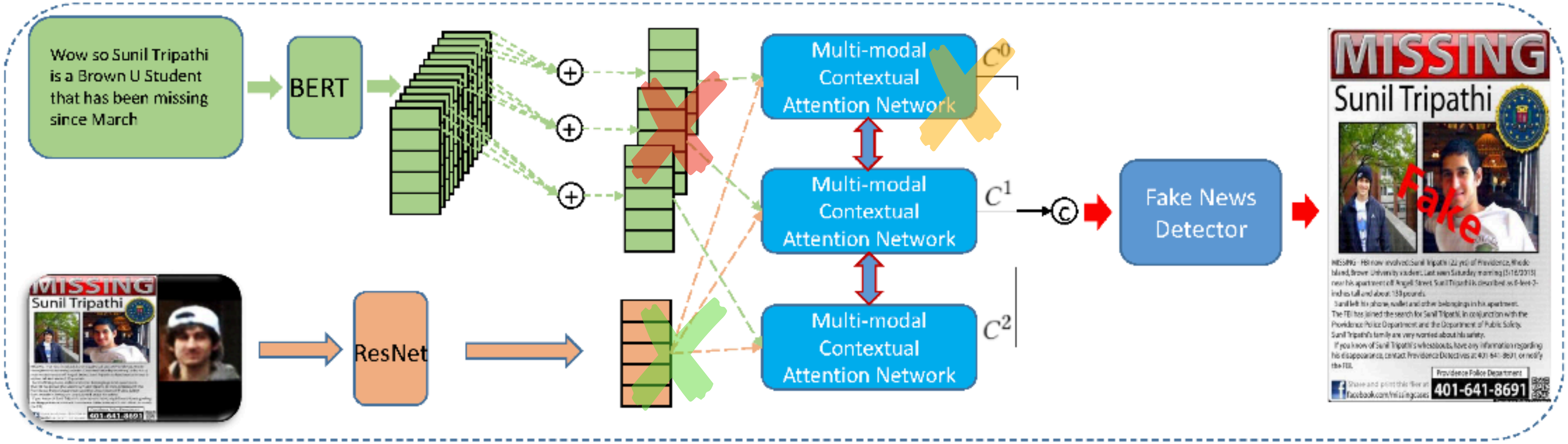
Experiments
Result & Analysis

- HMCAN outperforms all the baselines on Twitter and PHEME datasets.
- Observe that on Weibo dataset, in the case of fake news, the F1 and accuracy of HMCAN are lower than SpotFake*, while in the case of real news, the F1 of HMCAN is higher.
- Demonstrate that the proposed model can jointly model multi-modal context information and hierarchical semantics of text in a unified deep model.
  - which can better capture the underlying representation of posts, so as to improve the performance of fake news detection.

Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	att_RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SpotFake	0.869	0.877	0.859	0.868	0.861	0.879	0.870
	SpotFake*	**0.892**	0.902	0.964	**0.932**	0.847	0.656	0.739
	SpotFake+	0.870	0.887	0.849	0.868	0.855	0.892	0.873
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	**0.890**
TWITTER	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	att_RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SpotFake	0.771	0.784	0.744	0.764	0.769	0.807	0.787
	SpotFake*	0.777	0.751	0.900	0.820	0.832	0.606	0.701
	SpotFake+	0.790	0.793	0.827	0.810	0.786	0.747	0.766
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	SVM-TS	0.639	0.546	0.576	0.560	0.729	0.705	0.717
	GRU	0.832	0.782	0.712	0.745	0.855	0.896	0.865
	CNN	0.779	0.732	0.606	0.663	0.799	0.875	0.835
	SAFE	0.811	0.827	0.559	0.667	0.806	0.940	0.866
	att_RNN	0.850	0.791	0.749	0.770	0.876	0.899	0.888
	EANN	0.681	0.685	0.664	0.694	0.701	0.750	0.747
	TextGCN	0.828	0.775	0.735	0.737	0.827	0.828	0.828
	MVAE	0.852	0.806	0.719	0.760	0.871	0.917	0.893
	SpotFake	0.823	0.743	0.745	0.744	0.864	0.863	0.863
	SpotFake+	0.800	0.730	0.668	0.697	0.832	0.869	0.850
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**

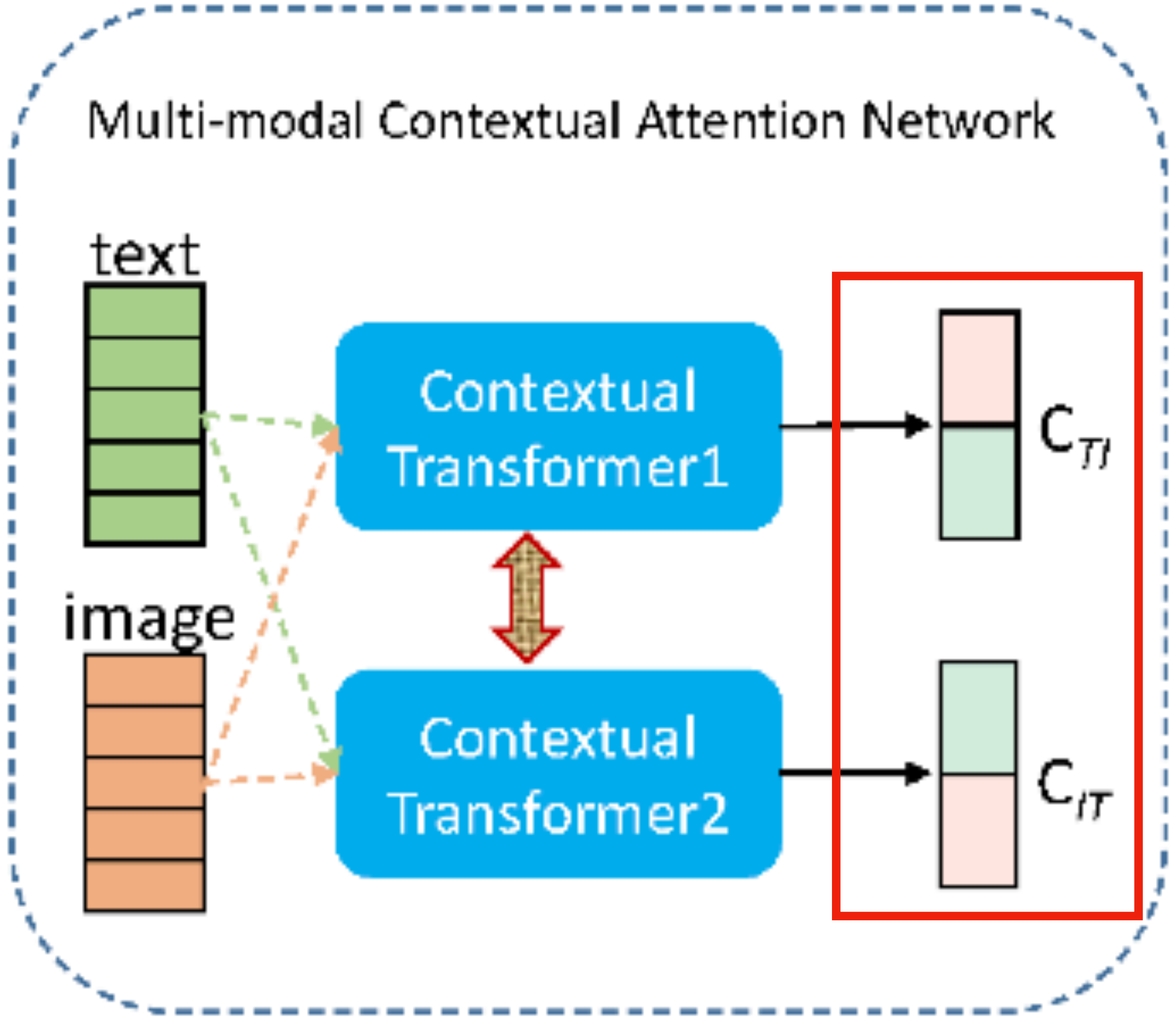
Experiments
Ablation Study (1/2)

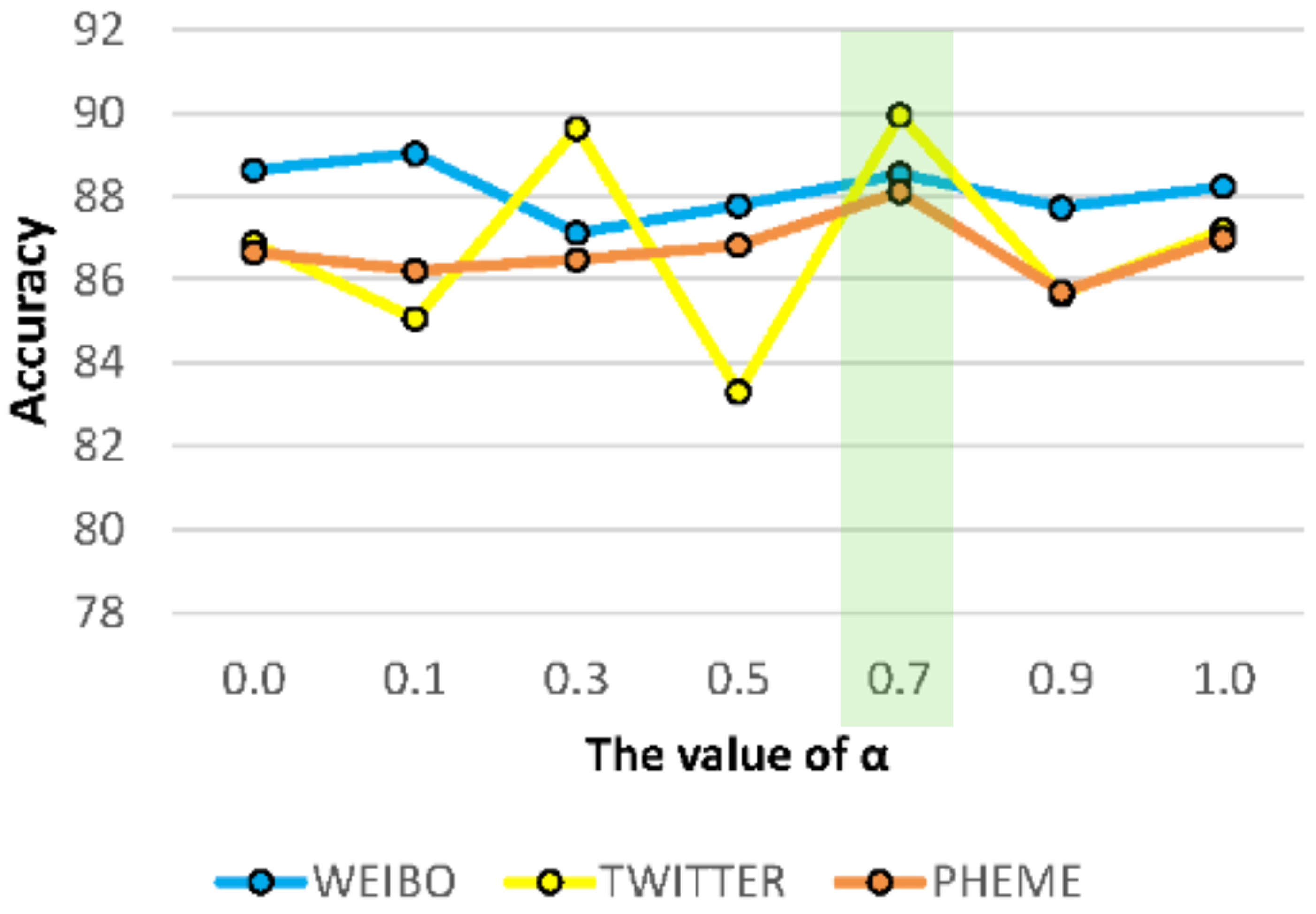
Dataset	Methods	Accuracy	Fake news			Real news		
			Precision	Recall	F1	Precision	Recall	F1
WEIBO	HMCAN¬*V*	0.809	0.832	0.774	0.802	0.788	0.843	0.815
	HMCAN¬*C*	0.872	0.902	0.836	0.868	0.847	0.909	0.877
	HMCAN¬*H*	0.877	0.871	0.885	0.878	0.883	0.869	0.876
	HMCAN	**0.885**	0.920	0.845	**0.881**	0.856	0.926	**0.890**
TWITTER	HMCAN¬*V*	0.755	0.828	0.590	0.689	0.719	0.896	0.798
	HMCAN¬*C*	0.790	0.886	0.622	0.731	0.743	0.932	0.827
	HMCAN¬*H*	0.879	0.884	0.849	0.866	0.875	0.906	0.890
	HMCAN	**0.897**	0.971	0.801	**0.878**	0.853	0.979	**0.912**
PHEME	HMCAN¬*V*	0.854	0.814	0.763	0.788	0.873	0.904	0.888
	HMCAN¬*C*	0.858	0.788	0.821	0.804	0.899	0.878	0.888
	HMCAN¬*H*	0.871	0.808	0.828	0.818	0.906	0.894	0.900
	HMCAN	**0.881**	0.830	0.838	**0.834**	0.910	0.905	**0.907**



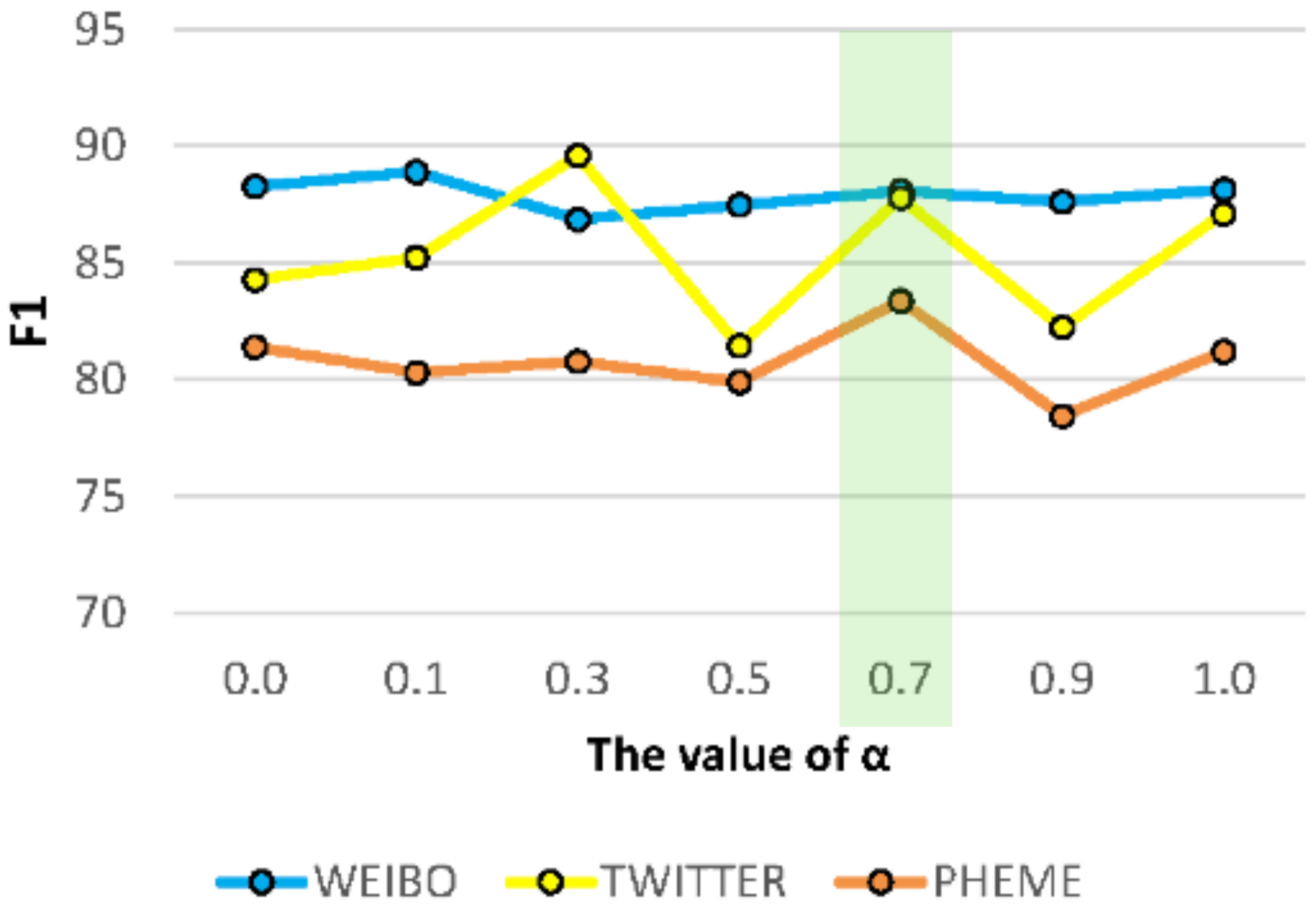
Experiments

Impact of the value α





(a) Accuracy

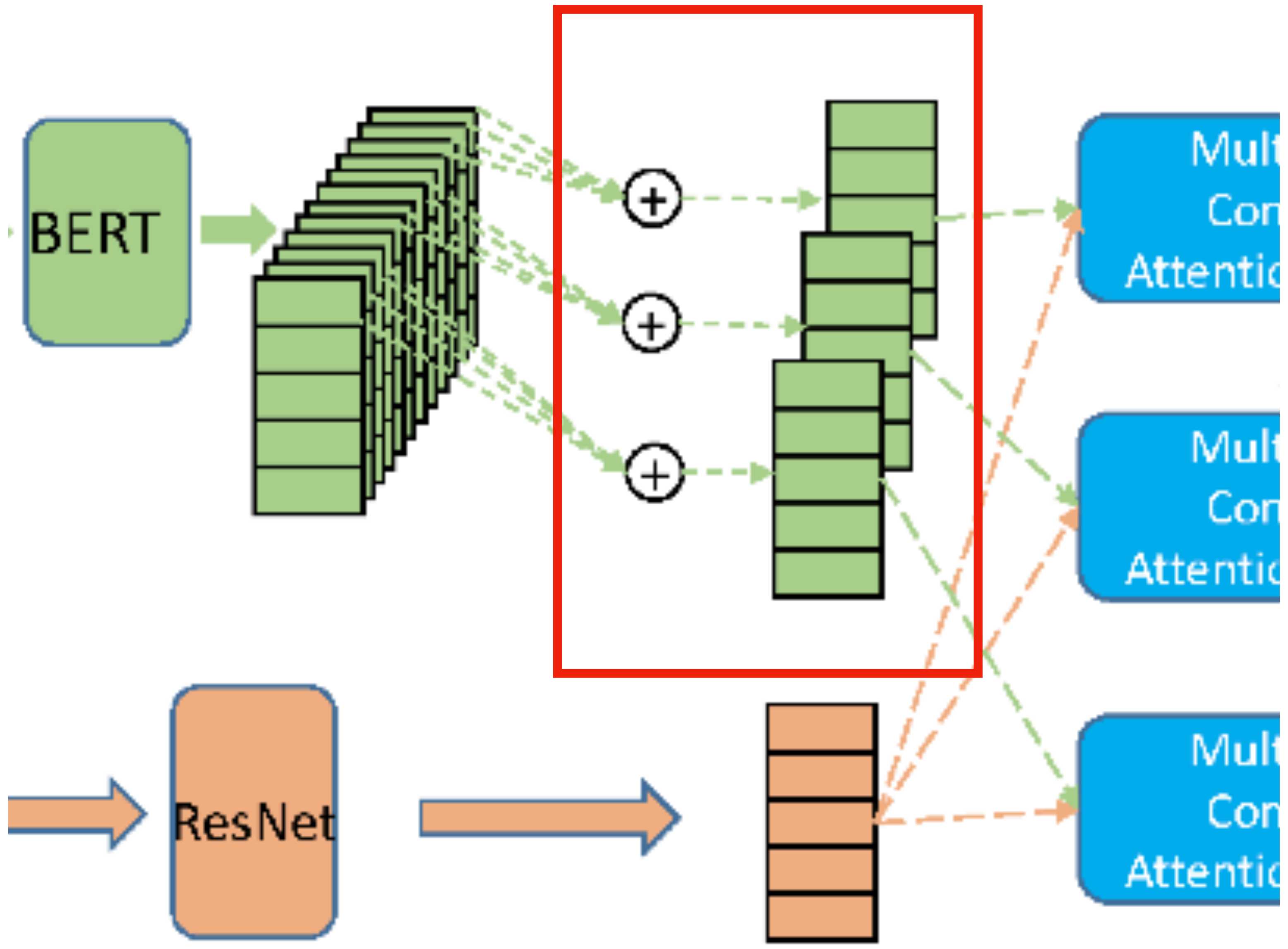


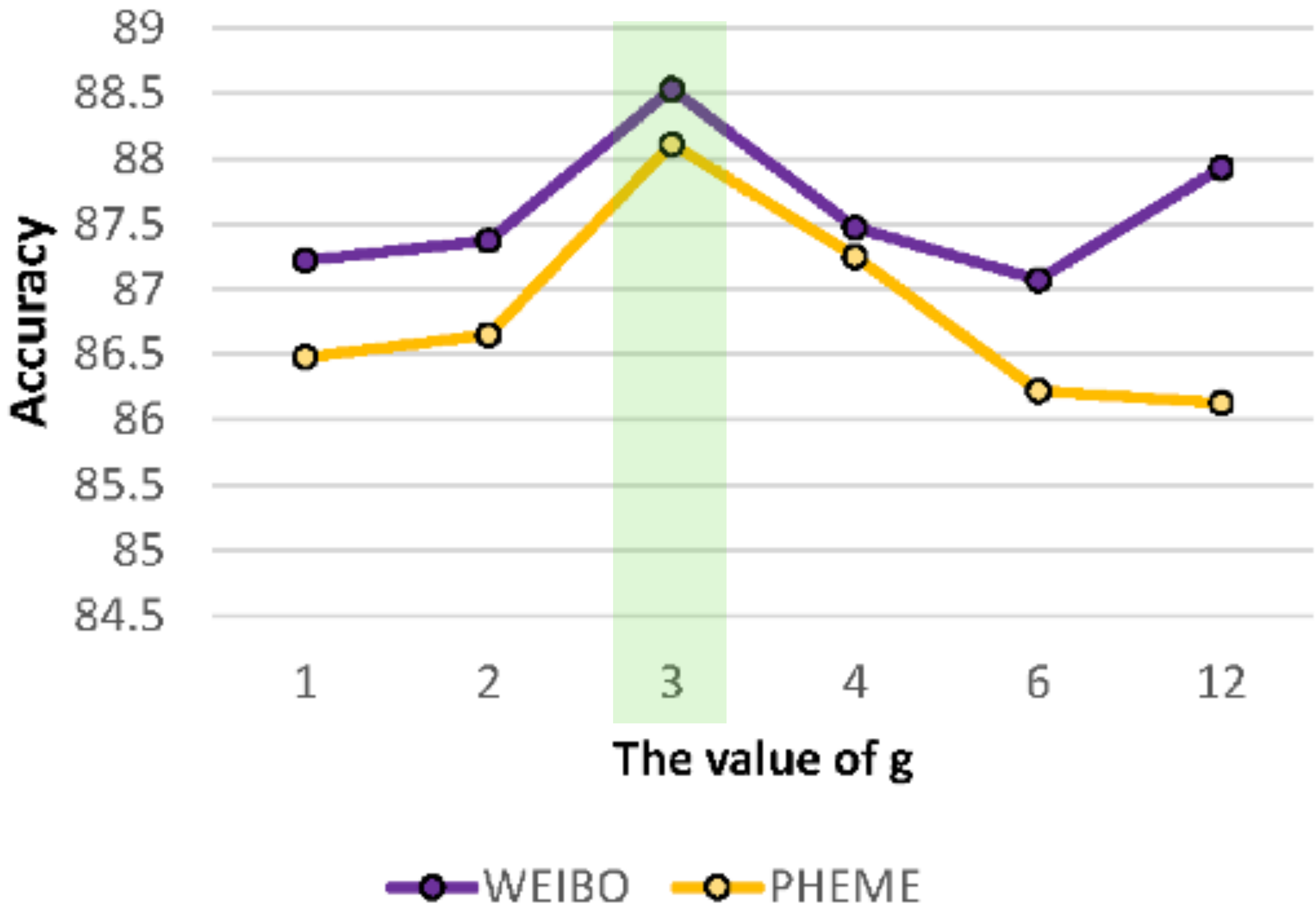
(b) F1 score of fake news

- $C = \alpha C_{TI} + \beta C_{IT}$
- $\alpha + \beta = 1$

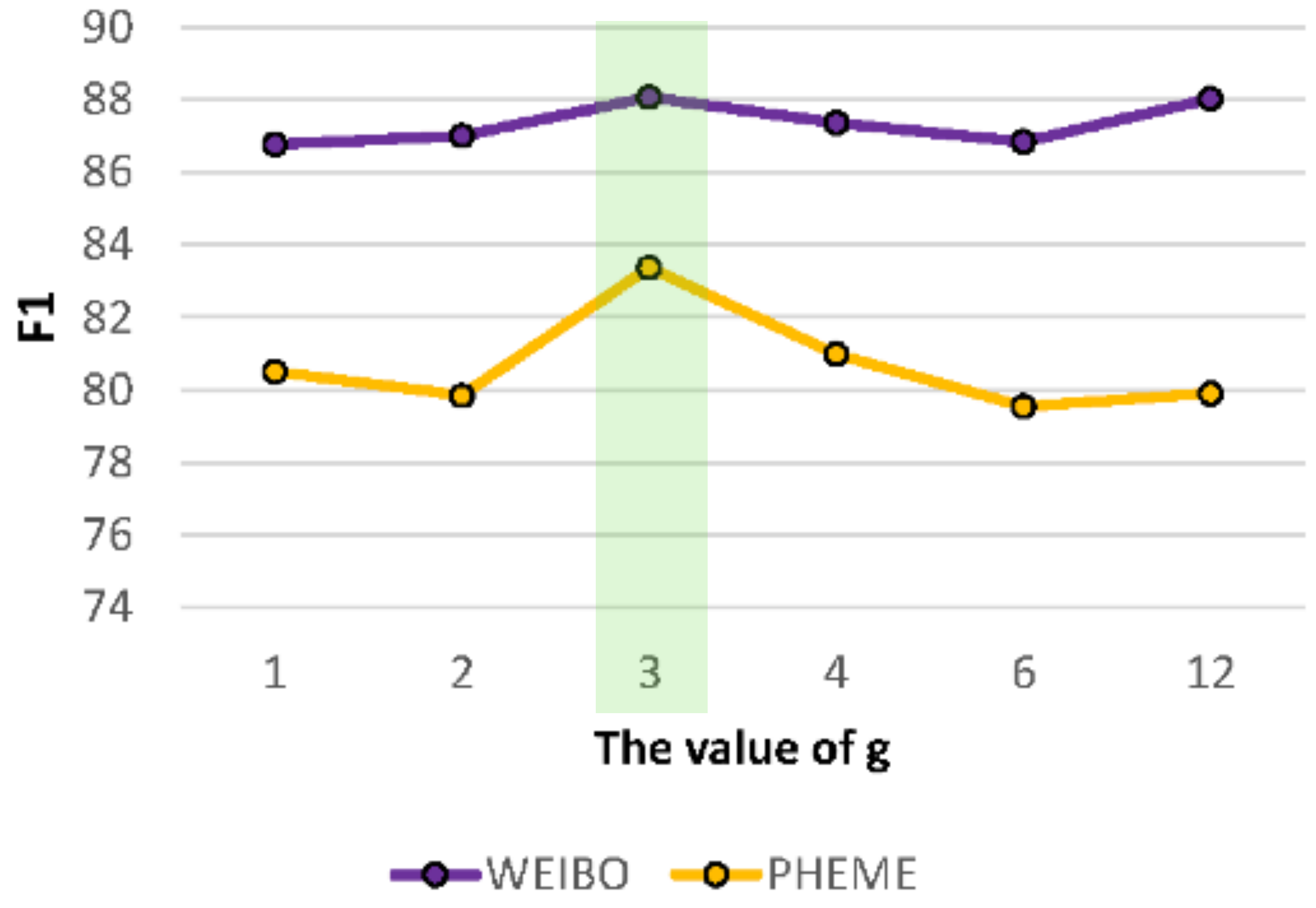
Experiments

Impact of the value g





(a) Accuracy



(b) F1 score of fake news

Conclusion
of HMCAN

- Propose a novel hierarchical multi-modal contextual attention network (HMCAN) for fake news detection task.
  - To jointly model the multi-modal context information and the hierarchical semantics of text in a unified deep model.
    - Employ ResNet to learn better representations of images and utilize BERT to embed the textual content of news.
    - A multi-modal contextual attention network is proposed to fuse both inter-modality and intra-modality relationships.
    - Design a hierarchical encoding network to capture the rich hierarchical semantics for fake news detection.

Comments
of HMCAN

- Simple but effective method.
  - Use BERT & ResNet as initial encoder.
  - Design a contextual attention network to get complementary information.
  - Utilize the advantage of BERT intermediate layer to get more information.
- Parameter not enough flexible and a little sensitive.