Multimodal Fusion with Co-Attention Networks for Fake News Detection

Yang Wu [1,2], Pengwei Zhan[1,2], Yunjian Zhang[1,2], Liming Wang[1]*, Zhen Xu[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
{wuyang0419,zhanpengwei,zhangyunjian,wangliming,xuzhen}@iie.ac.cn

ACL IJCNLP'21 (International Joint Conference on Natural Language Processing)

220808 Chia-Chun Ho

Outline
of MCAN

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction

Fake News Detection

- The rapid growth of social media has created fertile soil for the emergence and fast spread of fake news.
  - U.S. 2016 presidential election, COVID-19
- Tweets with images are getting popular on social media recently.
  - That have richer information and attract more viewers than tweets with only text.
  - Fake news also make full use of this advantage to draw and mislead readers.

Introduction

Examples of Fake News on Twitter

Sharks in the mall! After the hurricane sandy!

Lenticular Clouds over Mount Fuji.

Woman, 36, gives birth to 14 children from 14 different fathers.

Introduction

Existing Approaches (1/2)

- Some studies explore to learn the joint representations of text and image.
  - EANN, MAVE
    - Nevertheless, they are not fine-grained enough in feature extraction and feature fusion.
- Some studies require labor-intensive extra information.
  - Social context, event category

Introduction

Existing Approaches (2/2)

- Except for text in tweets, the methods mentioned above all focus on characteristics of images at the semantic level.
  - e.g., emotional provocations, which can be reflected in the spatial domain.
- Some models obtain fused representations by simply concatenating multi-modality features.
  - Although leverages local attention mechanism, still cannot reflect the similarity between textual-social representations and visual representations.

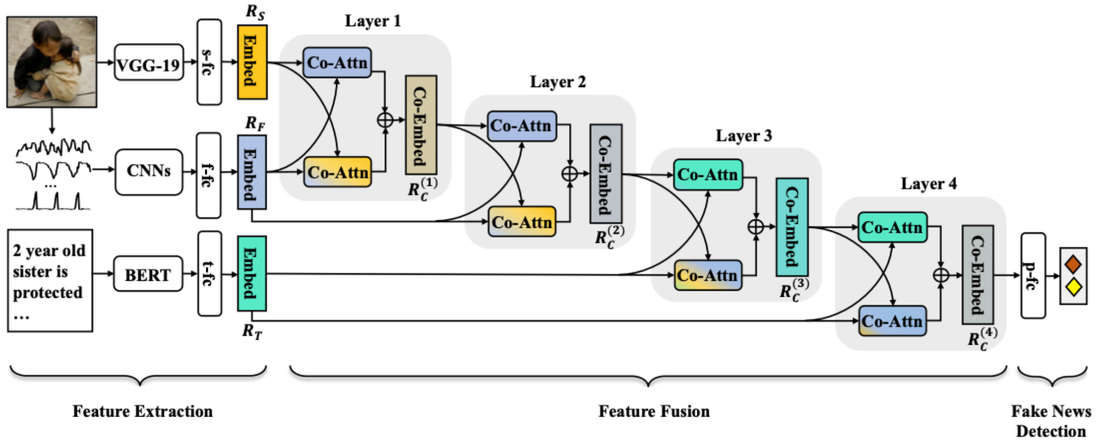
Introduction

Inter-modality Attention Relations

- Intuitively, when people judge news credibility with text and image, they often observe image first and then read text.
  - This process may be repeated several times.
  - In this process, people understand image according to the textual information, and understand text according to the associated image information.
- So the information of one modality is conditionally fused with that of another modality for once or multiple times.
- Intuitively, there are inter-modality attention relations between image and text.

Introduction

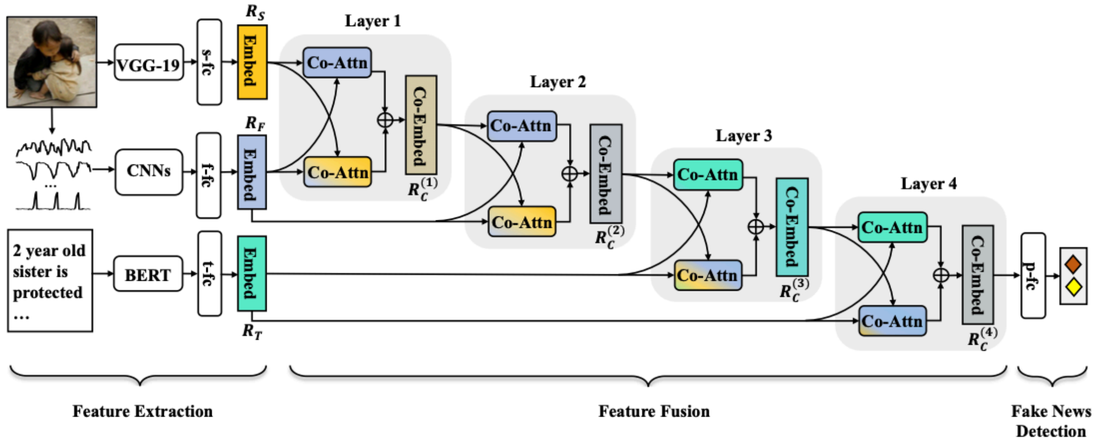
MCAN



- Propose the Multimodal Co-attention Networks for FND.
  - First extract spatial-domain features and frequency-domain features from image, as well as text features from text.
  - Then develop a novel fusion approach with multiple co-attention layers to learn inter-modality relations.
    - Fuses visual features first, and then the textual features.
  - The fused representation obtained from the last co-attention layer is used for FND.

Introduction

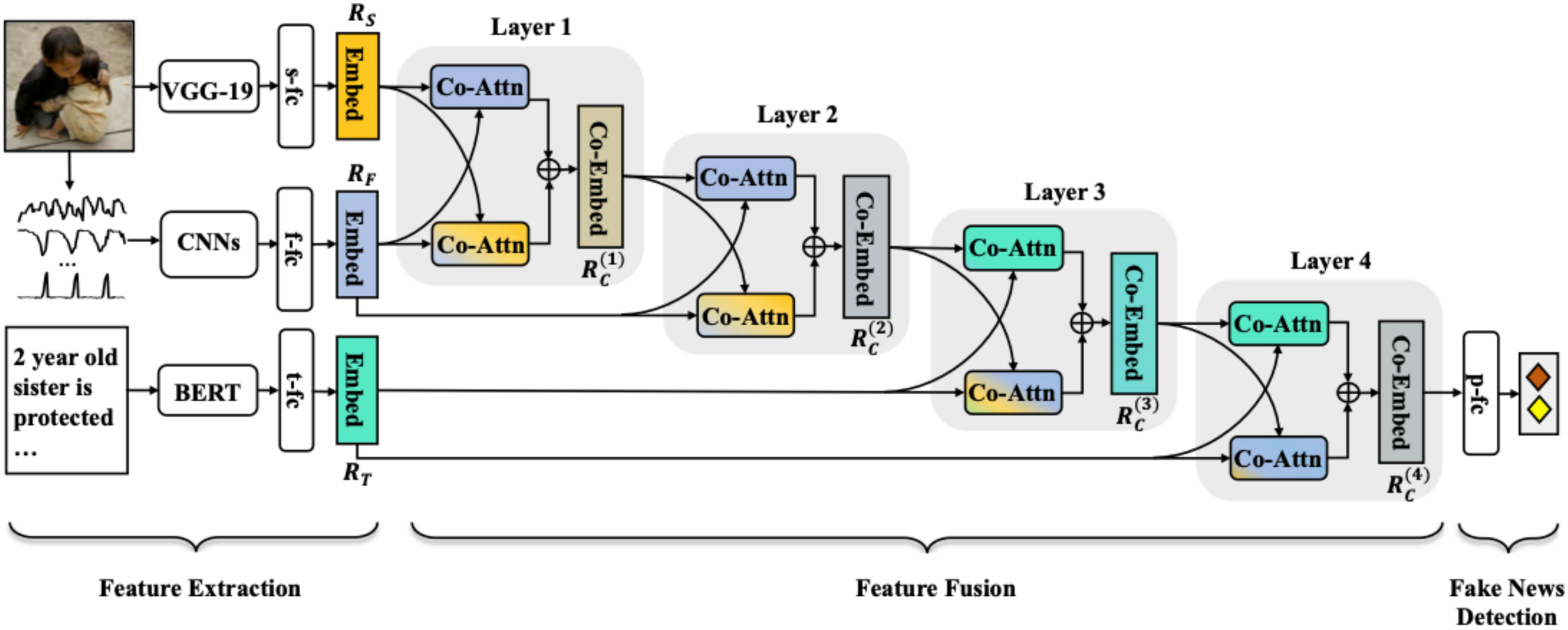
Contributions



- Propose a novel end-to-end approach to detect fake news on social media only using the text and the attached image.
  - Without any extra information and auxiliary tasks.
- Proposed MCAN model stacks multiple co-attention layers to fuse the multimodal features.
  - That can learn inter-dependencies among them.
- MCAN model is a general framework for fake news detection, and the components of MCAN are flexible.
  - The sub-networks used to extract multimodal features can be replaced by different models.

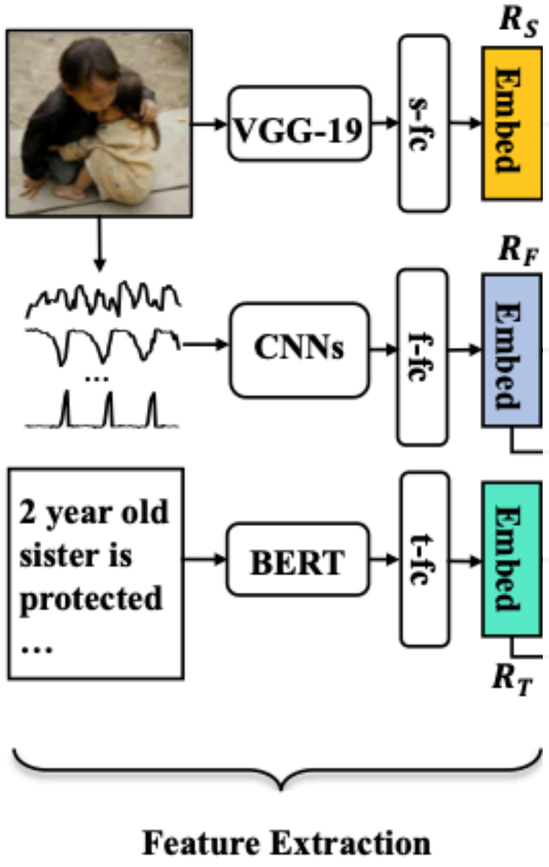
Methodology

Multimodal Co-attention Networks (MCAN)



Methodology

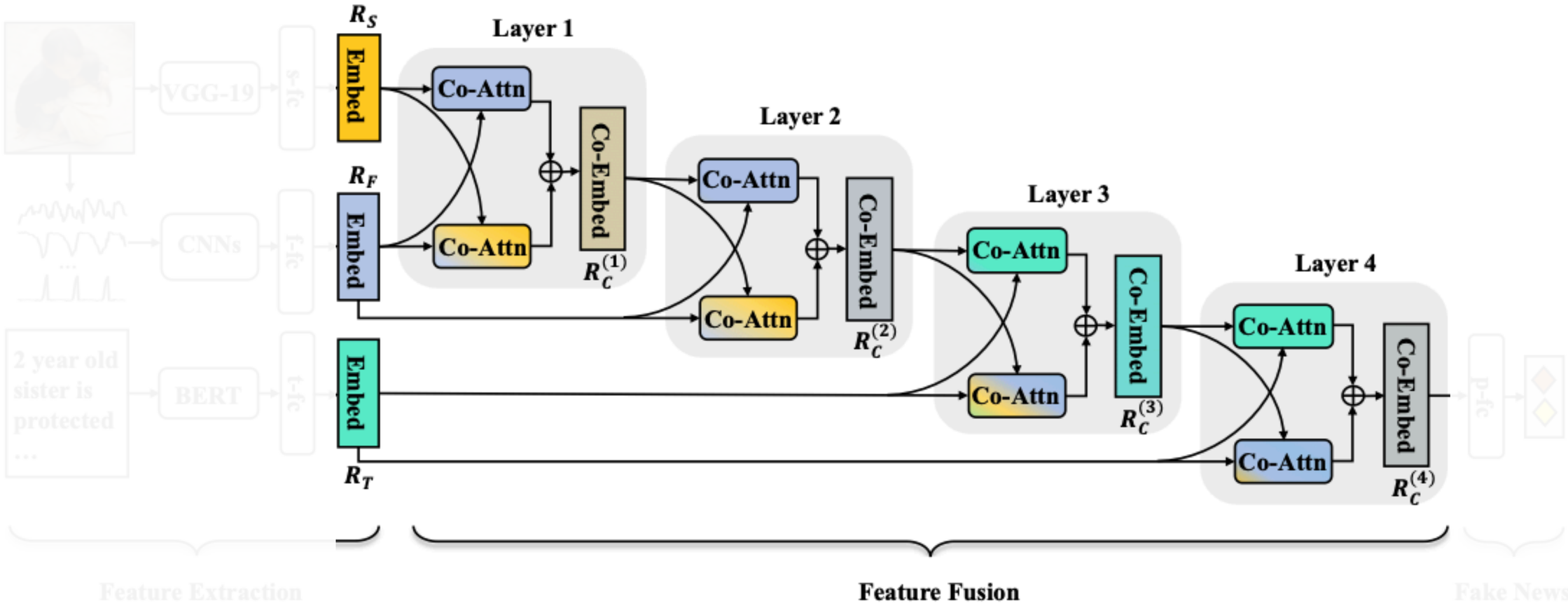
Feature Extraction



- Spatial-Domain Feature
  - VGG-19
- Frequency-Domain Feature
  - Design a CNN-based sub-network
- Textual Feature
  - BERT

Methodology

Feature Fusion



Methodology

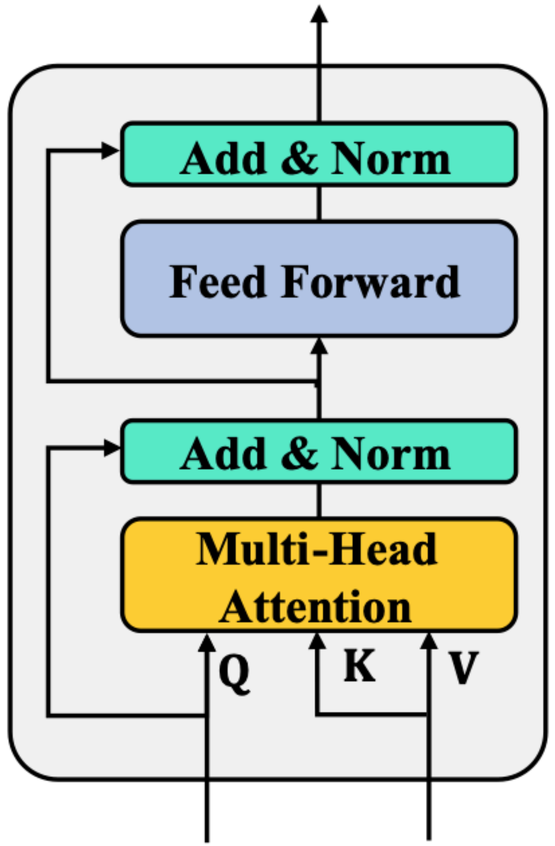
Feature Fusion

- Intuitively, people often look at the image first and then read the text when reading the news with image and text.
  - This process may be repeated several times, continuously fusing image and text information.
  - Therefore, develop a novel fusion approach to simulate this process.

Methodology

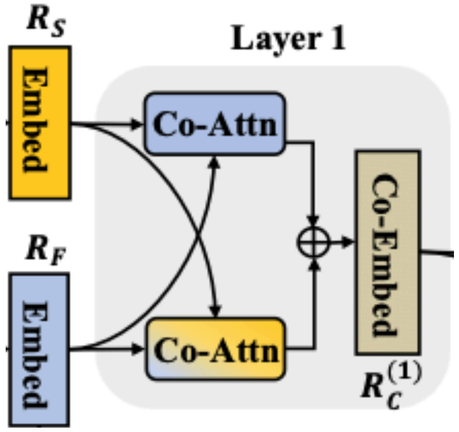
Co-attention Block (Co-Attn)

- The co-attention block is extended from the MSA block.
- The Co-Attn block produces an attention-pooled feature for one modality conditioned on another modality.
- If Q comes from text and K, V come from the attached image,
  - The attention value can be used as a measure of the similarity between the text and image, and then weights the image.
- Just like humans, after reading the text, they will pay more attention to the areas in the image that are similar to the text.



Methodology

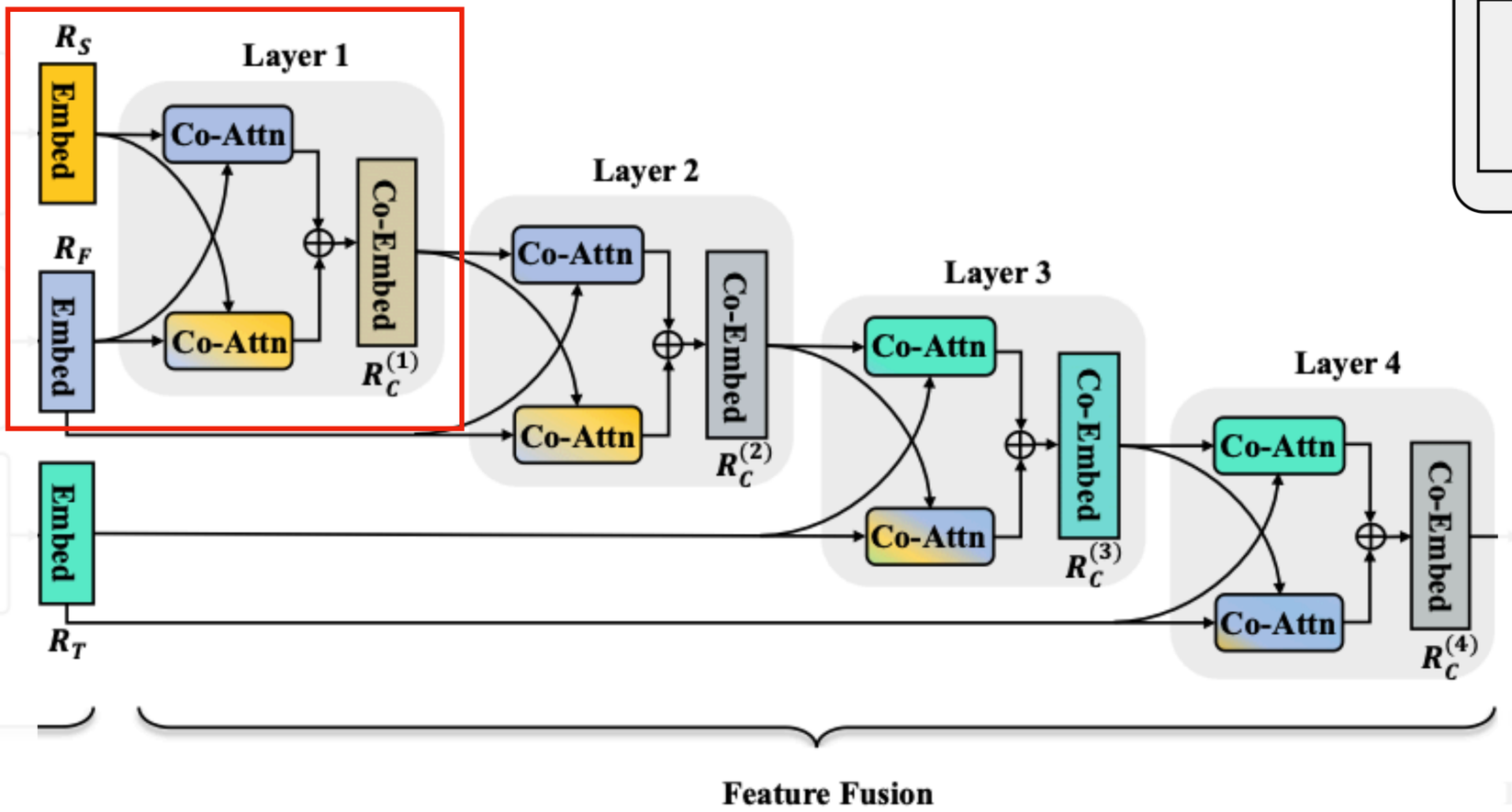
Co-attention Layer (CA Layer)

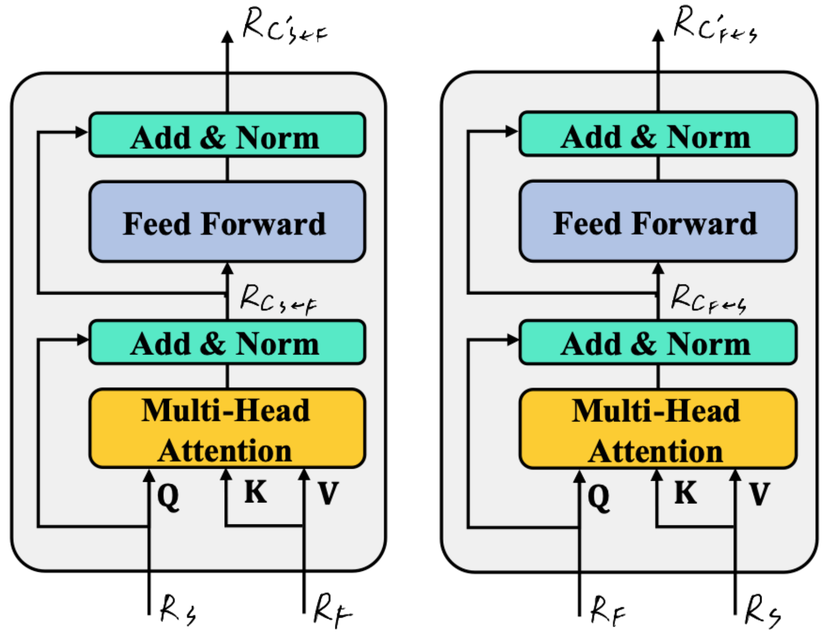


- Obtain a CA layer by connecting two Co-Attn blocks in parallel.
- Giving two Co-Attn blocks different features, the CA layer computes queries, keys, and values for each Co-Attn block.
  - Then the keys and values of one Co-Attn block are passed as input to another Co-Attn block.
- The outputs of two Co-Attn blocks are concatenated together and then fed into a fully connected layer to get the fused representation.
- The CA layer models dense interactions between input modalities by exchanging their information.

Methodology

Multiple Co-attention Stacking





$$R_{C_{S\leftarrow F}} = R_S + \mathrm{MA}(R_S, R_F, R_F)$$

$$R_{C'_{S\leftarrow F}} = R_{C_{S\leftarrow F}} + \mathrm{FFN}(R_{C_{S\leftarrow F}})$$

$$R_{C_{F\leftarrow S}} = R_F + \mathrm{MA}(R_F, R_S, R_S)$$

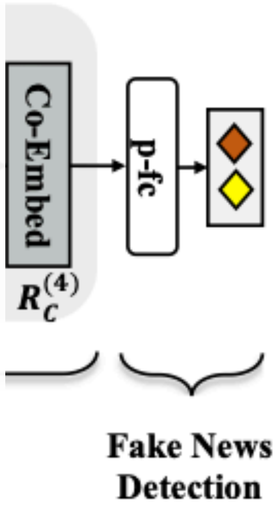
$$R_{C'_{F\leftarrow S}} = R_{C_{F\leftarrow S}} + \mathrm{FFN}(R_{C_{F\leftarrow S}})$$

$$R_C^{(1)} = (R_{C'_{S\leftarrow F}} \oplus R_{C'_{F\leftarrow S}})W_C^{(1)}$$

Methodology

Model Learning

- Obtained the multimodal feature representation $R_C^{(4)}$ fused features of text, spatial domain, and frequency domain.
- Let $f = R_C^{(4)}$, the output of the proposed MCAN is the probability of a tweet being fake:
  - $\hat{y} = \text{softmax}(\max(0, f W_f) W_s)$
- The loss function is devised to minimize the cross-entropy value:
  - $L(\Theta) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$



Experiments

Datasets

- Weibo
- Twitter

	Twitter	Weibo
# of fake news	8199	4211
# of real news	6681	3639
# of images	512	7850

Experiments
Baselines (1/2)

- Unimodal
  - Text, Spatial, Freq
- Multimodal
  - VQA: a model aims to answer questions according to the given images.
  - NeuralTalk: a deep recurrent framework for image caption.
  - att-RNN, EANN, MVAE
  - MCAN-A: without the part of fusing multimodal features. Spatial-domain features, frequency-domain features, and textual features are simply concatenated for prediction.

Experiments

Result & Analysis

- MCAN-A performs better than unimodal models.
  - Indicates that adding features usually improves model performance.
    - But it is not always positively correlated.
    - Text on Weibo dataset is better than MCAN-A.
- After adding the process of multimodal fusion, MCAN beats MCAN-A and other multimodal models.
  - Embodies proposed feature fusion method is indeed better than the simple concatenation method.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Text	0.633	0.656	0.762	0.705	0.587	0.459	0.515
	Spatial	0.671	0.841	0.527	0.648	0.574	0.864	0.69
	Freq	0.665	0.733	0.656	0.692	0.592	0.677	0.631
	VQA	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.81	0.498	0.617	0.584	0.759	0.66
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.73
	MCAN-A	0.737	0.840	0.671	0.746	0.65	0.827	0.727
	MCAN	**0.809**	**0.889**	**0.765**	**0.822**	**0.732**	**0.871**	**0.795**
Weibo	Text	0.876	0.885	0.871	0.878	0.865	0.878	0.871
	Spatial	0.857	0.85	0.877	0.863	0.863	0.834	0.848
	Freq	0.717	0.728	0.724	0.726	0.706	0.710	0.708
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	NeuralTalk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.72	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MCAN-A	0.869	0.868	0.879	0.874	0.869	0.857	0.863
	MCAN	**0.899**	**0.913**	**0.889**	**0.901**	**0.884**	**0.909**	**0.897**

Experiments

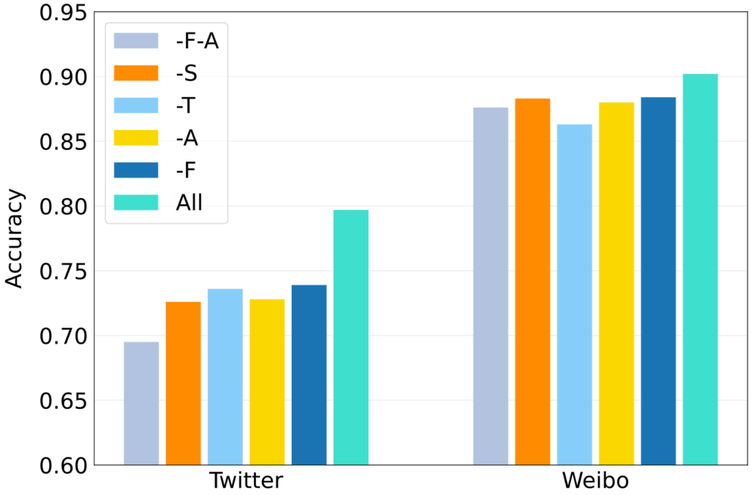
Result & Analysis

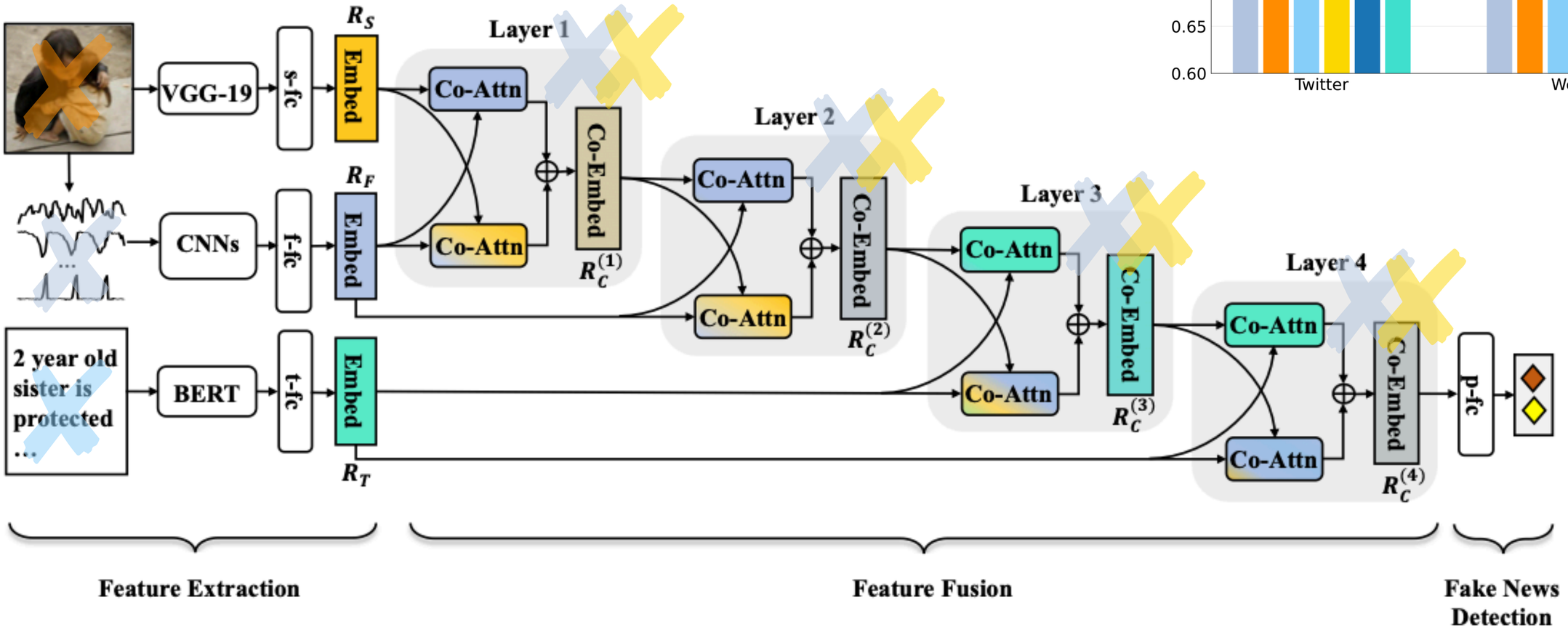
- On Weibo dataset, the accuracy of fine-tuned BERT and VGG-19 all exceed 85%.
  - In this case, MCAN further improves the accuracy to close to 90% with the help of cascaded way of stacking CA layers.
  - Comparing with the situation on Twitter dataset, can find that MCAN performs better in the face of weak unimodal features.
- In MCAN model, the representation ability of features can be greatly improved by effectively fusing other features.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
Twitter	Text	0.633	0.656	0.762	0.705	0.587	0.459	0.515
	Spatial	0.671	0.841	0.527	0.648	0.574	0.864	0.69
	Freq	0.665	0.733	0.656	0.692	0.592	0.677	0.631
	VQA	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	NeuralTalk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.81	0.498	0.617	0.584	0.759	0.66
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.73
	MCAN-A	0.737	0.840	0.671	0.746	0.65	0.827	0.727
	MCAN	**0.809**	**0.889**	**0.765**	**0.822**	**0.732**	**0.871**	**0.795**
Weibo	Text	0.876	0.885	0.871	0.878	0.865	0.878	0.871
	Spatial	0.857	0.85	0.877	0.863	0.863	0.834	0.848
	Freq	0.717	0.728	0.724	0.726	0.706	0.710	0.708
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	NeuralTalk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.854	0.656	0.742	0.72	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MCAN-A	0.869	0.868	0.879	0.874	0.869	0.857	0.863
	MCAN	**0.899**	**0.913**	**0.889**	**0.901**	**0.884**	**0.909**	**0.897**

Experiments

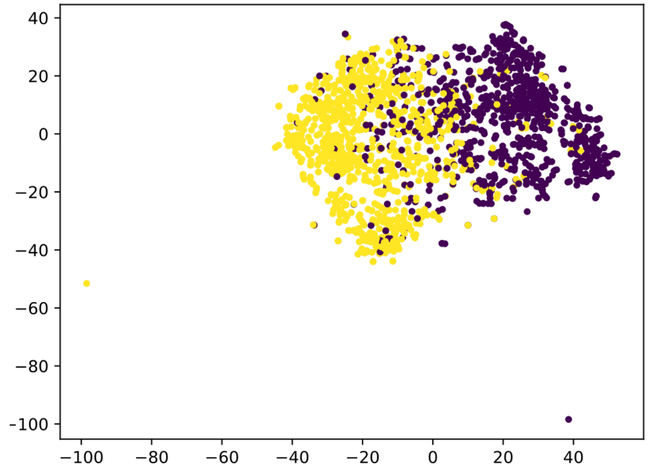
Quantitative Analysis (1/2)



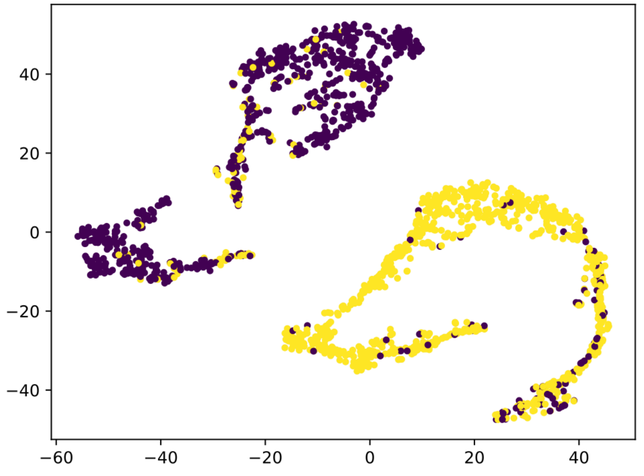


Experiments

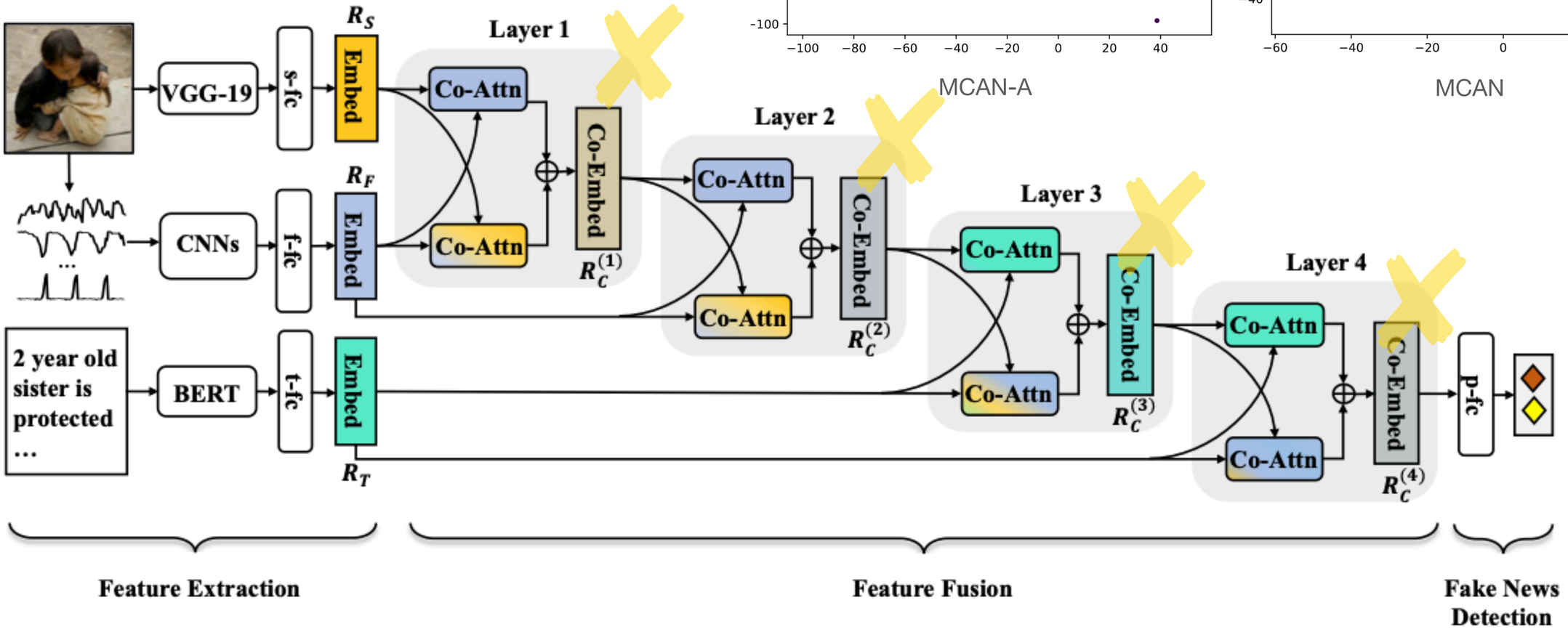
Quantitative Analysis (2/2)



MCAN-A



MCAN



Conclusion
of MCAN

- Propose a novel Multimodal CoAttention Networks (MCAN) to tackle the challenge of fusing multimodal (textual and visual) features for fake news detection.
- Utilize three different sub-networks to extract features from text, spatial domain, and frequency domain, respectively.
- Then the three features are deeply fused by stacking co-attention layers, which is inspired by human behavior.
    - When people read news with image, image and text are read once or multiple times, and continuously fused in brain.

Comments
of MCAN

- Inspired from human reading behavior to design the stack co-attention layer.
  - To simulate the view image first then read the text.
- Baselines in this paper all before from 2019 too old to prove the effective.