Leveraging Intra and Inter Modality Relationship for Multimodal Fake News Detection

Shivangi Singhal
IIIT Delhi
India
shivangis@iiitd.ac.in

Tanisha Pandey
IIIT Delhi
India
tanisha17116@iiitd.ac.in

Saksham Mrig
IIIT Delhi
India
saksham19385@iiitd.ac.in

Rajiv Ratn Shah
IIIT Delhi
India
rajivratn@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT Hyderabad
India
pk.guru@iiit.ac.in

WWW'22

220912 Chia-Chun Ho

Outline
of LIIRM

Introduction

Problem Formulation

Methodology

Experiments

Conclusion

Comments

Introduction

News Example

- *A husband divided his assets in half while settling the divorce case with her ex-wife.*

- At a first read, the text might look believable.

- Now, when we read the same piece of information but with an image, shown in Figure 1, we might question the credibility of news.

Figure 1: An example of the tweet from the Twitter Dataset [4]. The corresponding text reads, 'Husband Gave His Unfaithful Ex-Wife Half Of Everything He Owned – Literally'. Our proposed intra-modality feature extractor curates the fine-grained salient representations for image and text, represented in the *blue* and *red* color, respectively.

Introduction

Existing Approaches

- A modality is strong when it can assign a high probability to the correct class.
- A higher probability implies a more informative signal and stronger confidence.
- Existing methods for multimodal FND do not work on the principles of weak and strong modality.
    - Instead, methods capture high-level information from different modalities and jointly model them to determine the authenticity of news.
    - The feature extraction also occurs globally, ignoring the salient pixels containing meaningful information.

Introduction

News Example

- For instance, Figure 1 highlights essential segments of the image and text containing details.
- However, current method of extracting visual features includes background information that might be unwanted.
- Similarly, there is a need to extract contextual dependencies for the textual features.

Figure 1: An example of the tweet from the Twitter Dataset [4]. The corresponding text reads, 'Husband Gave His Unfaithful Ex-Wife Half Of Everything He Owned – Literally'. Our proposed intra-modality feature extractor curates the fine-grained salient representations for image and text, represented in the *blue* and *red* color, respectively.

Introduction

Importance of Modalities

- Hypothesize that not all modalities play an equal role in the decision-making process on any particular sample.
  - Aim to design an architecture that utilizes a multiplicative multimodal method to capture inter-modality relationship.
  - The method suppresses the cost of a weaker modality by introducing a down-weight factor in the cross-entropy loss function.
    - The down-weight factor associated with each modality highlights the average prediction power of the remaining modalities.

Introduction
Intra-Relationship

- Capture the intra-modality relationship.
  - The idea is to generate fragments of a modality and then learn fine-grained salient representations from the fragments.
    - For image modality, perform bottom-up attention to extract the image patches.
    - The complex relationship between the patches is then encoded via self-attention mechanism.
    - The final visual representation is obtained by performing an average pooling operation over the fragment representations, resembling bag-of-visual-words model.

Introduction

Intra-Relationship

- Capture the intra-modality relationship.
  - The idea is to generate fragments of a modality and then learn fine-grained salient representations from the fragments.
    - For text modality, use a wordpiece tokenizer to generate text fragments, then using BERT to extract contextual representations.
    - The obtained embeddings are further passed through 1D-CNN to extract the phrase-level information.
    - The resultant text representation is obtained by passing intermediate learned representations via a fully connected layer.

Introduction

Contributions

- Capturing inter-modality relationship
  - Present a novel architecture that uses a multiplicative multimodal method to capture the inter-modality relationship between modalities.
  - Using the multiplicative multimodal method, aim to leverage information from a more reliable modality than a less reliable one on a per-sample basis.
- Capturing intra-modality relationship
  - The proposed method captures intra-modality relationship by extracting the fine-grained salient representations for image and text.
  - The resultant feature vectors capture rich contextual dependencies present within its components.

Problem Formulation

Notations

- A set of n news articles, $S = \{S_i^T, S_i^I, y\}_{i=1}^n$
  - S_i^T: text content, S_i^I: corresponding image, y: label (fake: $y = 0$, true: $y = 1$)
  - Every content piece comprises of k sentences: $\{S_i^{T_a}\}_{a=1}^k$
    - Each sentence $S_i^{T_a}$ is further tokenized into $\{w_{i_1}, w_{i_2}, \ldots, w_{i_k}\}$
  - Every image is segregated into a finite set of $\{m_i^1, m_i^2, \ldots, m_i^{36}\}$ fragments via bottom up attention module.

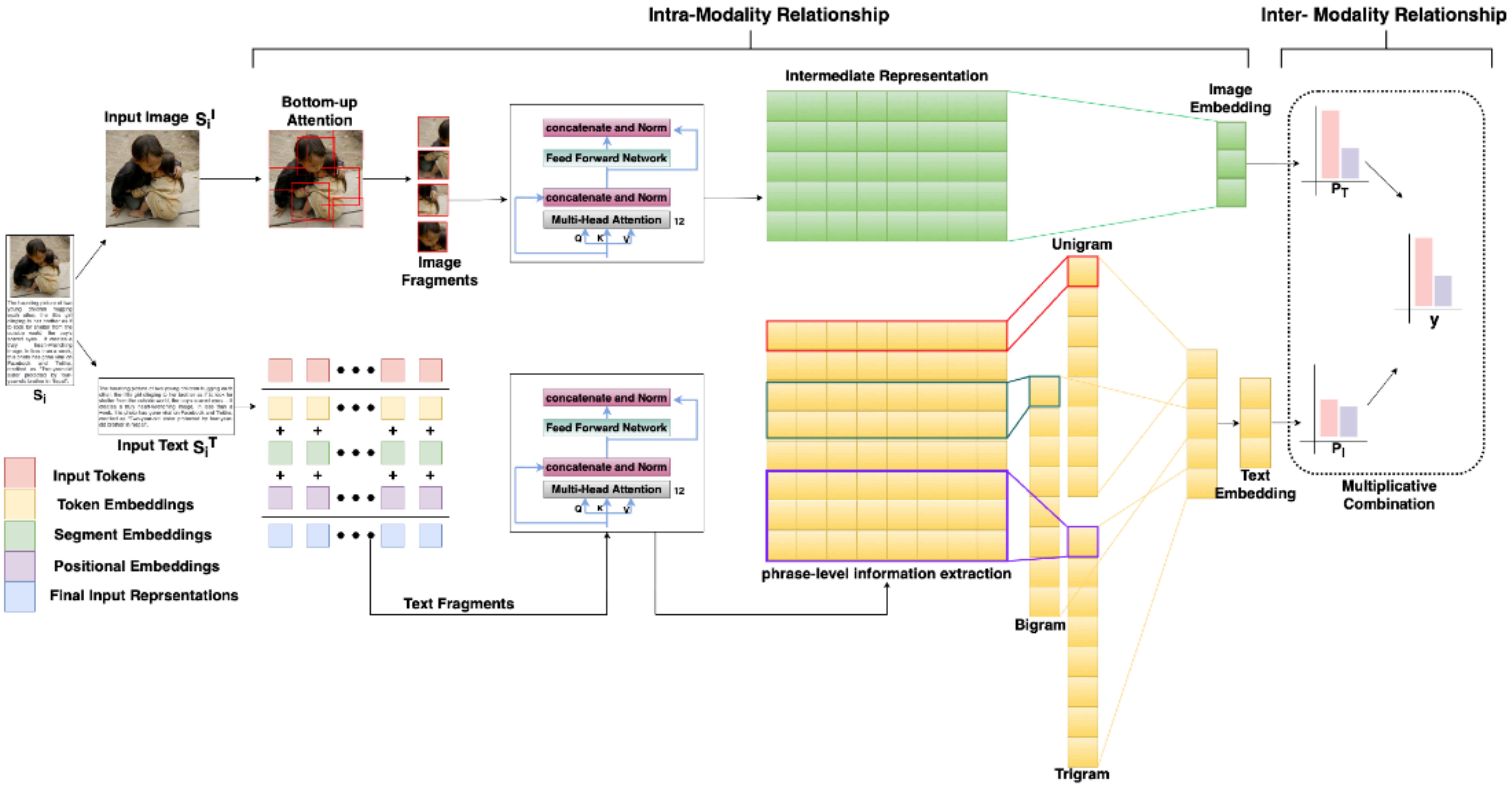
Problem Formulation

Problem

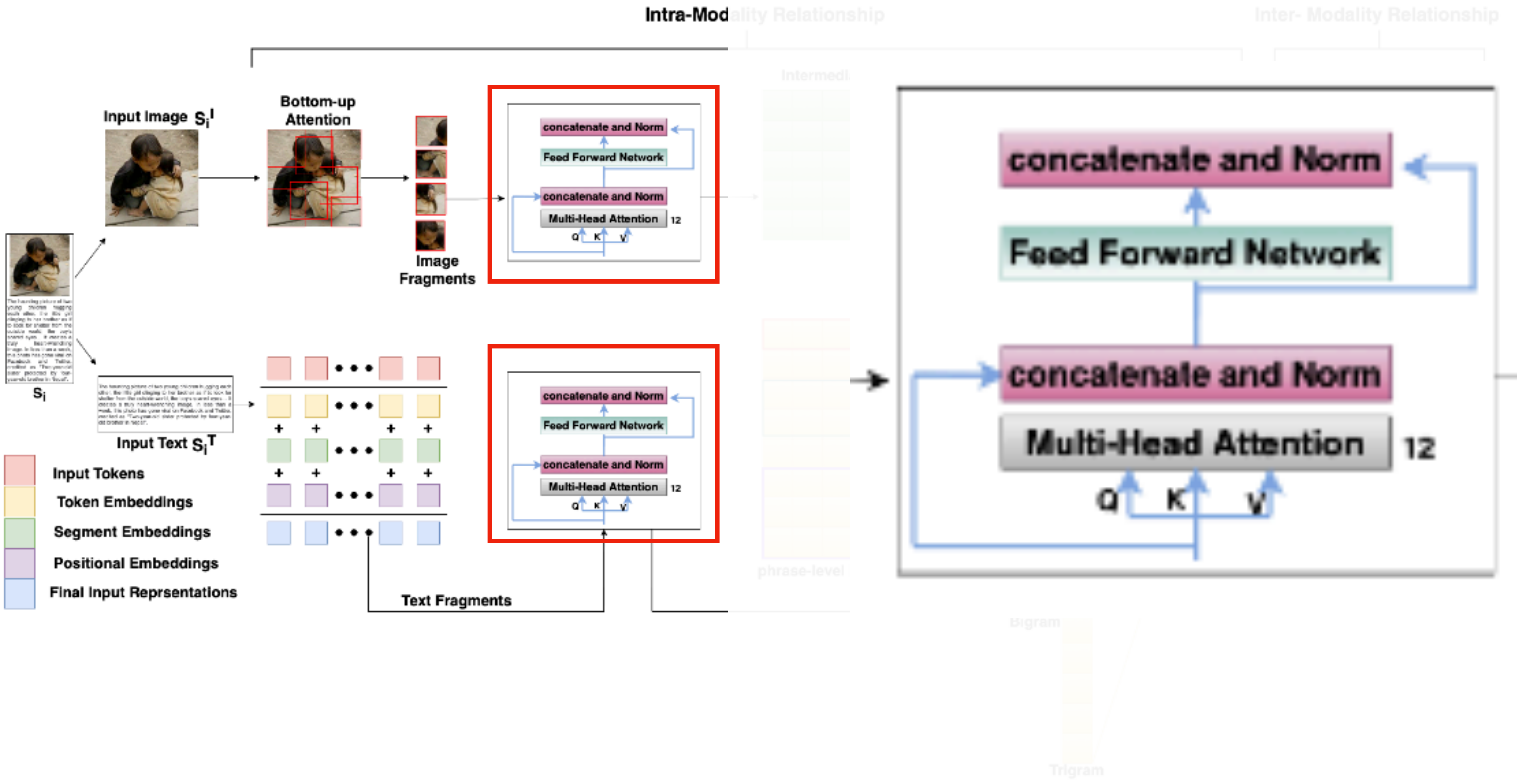
- Give a news sample $S = \{S_i^T, S_i^I, Y\}$.
- The goal is to design a novel architecture that capture
  - The intra-modality relationship via granular fragment representation &
  - Extracts the inter-modality relationship by inducing knowledge in classification sub-module that tell which modality contributed towards fakeness.
  - Such knowledge will also help readers understand the modality that contributed to the forgery.

Methodology

Proposed model

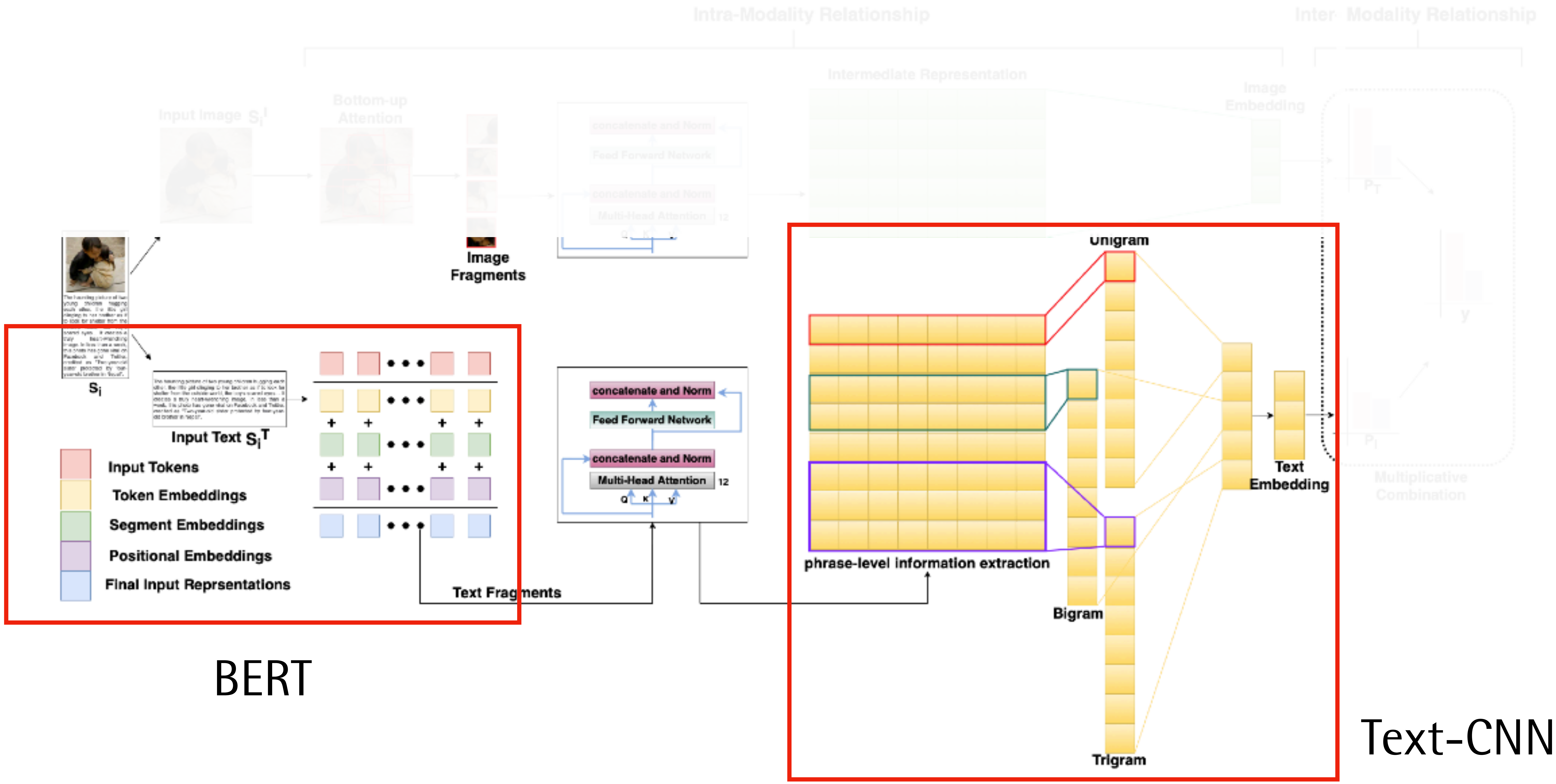


Methodology
Self-Attention



Methodology

Text-Embeddings

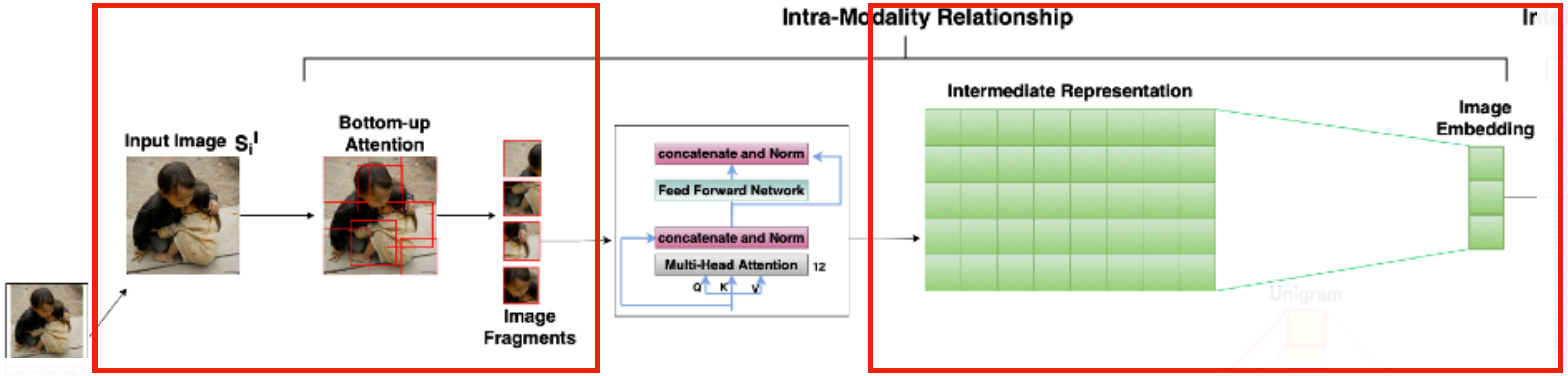


BERT

Text-CNN

Methodology

Image-Embeddings



Employ bottom-up attention model pre-trained on Visual Genome to extract a fixed-sized set of l patches.

The obtained image embeddings are condensed into a dense representations by performing average pooling followed by L2 normalization to procure the resultant image feature vector.

Methodology

Multiplicative Multimodal Method

- This work aims to capture interaction among different modalities to better perform the task at hand.
- There are some practical constraints in integrating synergies across modalities using existing additive approaches.
  - Additive methods assume that every modality is potentially helpful and is jointly combined to decide.
  - Neural network models built on top of aggregated features cannot determine the quality of each modality and its contribution toward fake detection tasks on a per-sample basis.

Methodology

Multiplicative Multimodal Method

- Given multiple input modalities, an ideal algorithm should be robust to noise from weak modalities and harvest relevant details from stronger modalities on a per sample basis.
- In this work, perform the multiplicative multimodal method that addresses the challenges mentioned above.
- Specifically, technique explicitly models that not all modalities contribute equally to any particular sample.

Methodology

Multiplicative Multimodal Method

- Let every modality present in a news sample make its own independent decision
- $P_T = [p^1, p^0]$, $P_I = [p^1, p^0]$, where P_T, P_I denotes the text and image predictions.
- Typical, additive combination would have resulted in
  - $l^y_{cross\_entropy} = -\sum_{i=1}^{M} \log\left(p_i^y\right)$
  - where l^y is a class loss as it is part of the loss function associated with a particular class.

Methodology

Multiplicative Multimodal Method

- To mitigate the challenges, utilized a down-weight scaling factor,
  - $$q_i = \left[\prod_{j \neq i} \left(1 - p_j \right) \right]^{\beta/(m-1)}$$
  - where β is a hyper-parameter used to control the strength of down-weighting.

Methodology

Down-weight factor

- The down-weight factor is responsible for suppressing the modality predictive power that incorrectly classifies the sample.
- For instance, if p_i show confident predictions for the correct class, down-weight factor will be a small value, suppressing cost for the other modalities $(j \neq i)$.
- Intuitively, when current modality gives a favourable prediction, other modalities need not be equally helpful.
- Larger the value of down-weight factor, stronger the suppressing effect on that modality and vice versa.

Methodology

Loss function

- Leverage benefits of extracting complementary information from the given piece of information using multiplicative method that have resulted in the modification of loss function as
- $l^y_{multiplicative} = -\sum_{i=1}^{M} q_i \cdot \log\left(p_i^y\right)$

Experiments

Datasets

- MediaEval
  - Train 7032: 5008 fake: real
  - Test 2564: 1217 fake: real
- Weibo
  - Train: Test 8:2
  - 4749 fake: 4779 real

Experiments

Baselines

- Single-modal
  - Text-CNN, BERT, VGG-19
- Multi-modal
  - EANN, MVAE, SpotFake

Experiments

Research Questions

- Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?
- How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection?
- Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample?

Experiments

Research Questions

- Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?
- How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection?
- Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample?

Experiments

Performance Analysis

- Results shown in the tables indicate that proposed method outperforms the baselines on accuracy and F1-score for Twitter and Weibo, respectively.
- SpotFake is the strongest baseline on multimodal fake news detection, and our proposed method outperforms it by a fair margin of an average of 3.05% and 4.525% on the accuracy and F1-score, respectively.

	MediaEval Benchmark Dataset			
Baselines	Acc	Prec.	Rec.	F1
Text-CNN†	0.614	0.599	0.612	0.594
BERT†	0.607	0.595	0.601	0.594
VGG-19$^{\mp}$	0.558	0.572	0.573	0.558
EANN‡	0.648	0.697	0.630	0.634
MVAE‡	0.745	0.745	0.748	0.744
SpotFake‡	0.777	0.791	0.753	0.760
Proposed	**0.831**	**0.836**	**0.832**	**0.830**

	Weibo Dataset			
Baselines	Acc	Prec.	Rec.	F1
Text-CNN†	0.794	0.791	0.800	0.792
BERT†	0.861	0.860	0.870	0.859
VGG-19$^{\mp}$	0.654	0.502	0.502	0.501
EANN‡	0.782	0.790	0.780	0.778
MVAE‡	0.824	0.830	0.822	0.823
SpotFake‡	0.8923	0.874	0.810	0.835
Proposed	**0.900**	**0.882**	**0.823**	**0.847**

Experiments

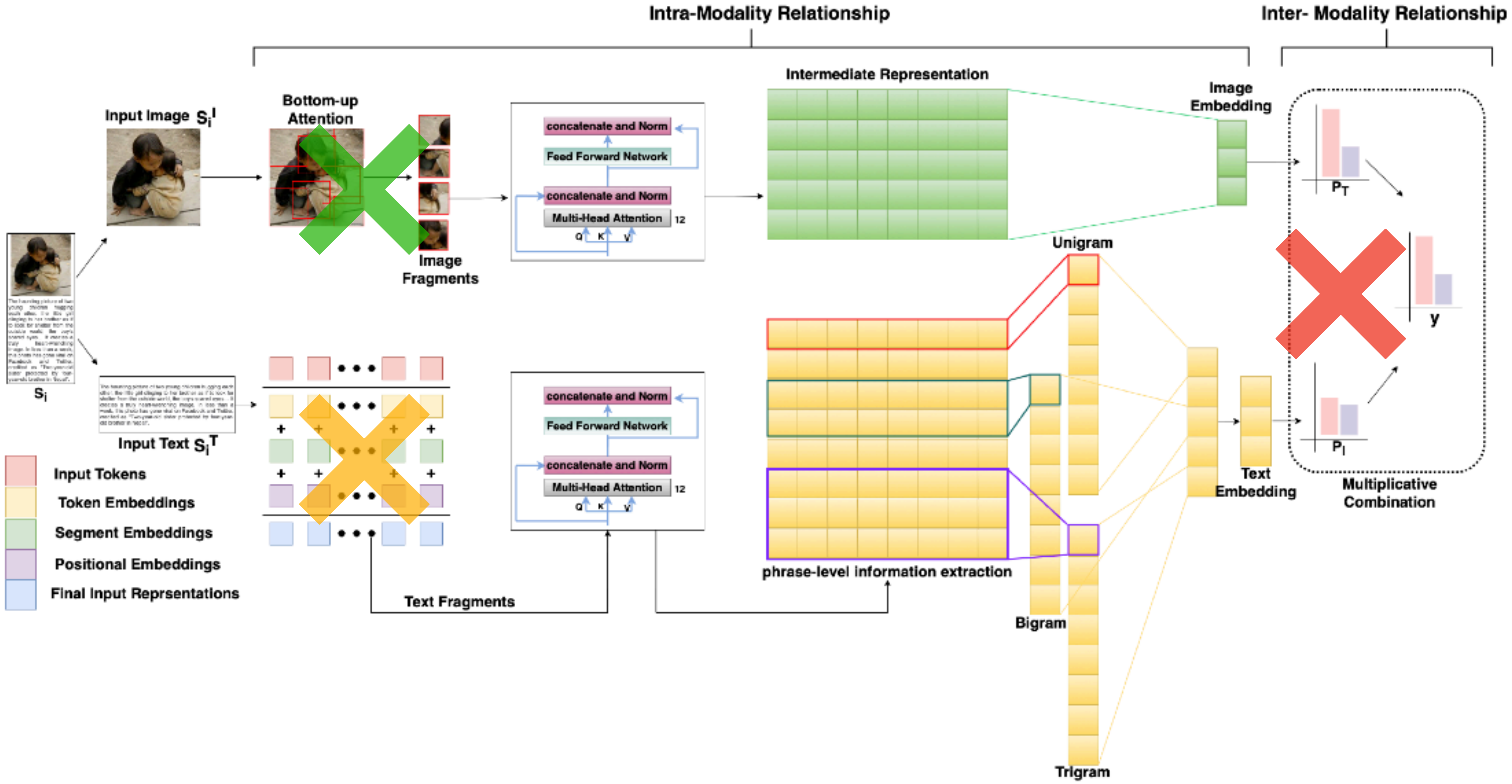
Research Questions

- Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?
- How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection?
- Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample?

Experiments

Ablation Analysis

	Variants	w/o Text	w/o Image	w/o Multiplicative	Proposed
Twitter	Acc	0.703	0.626	0.813	**0.831**
	Prec.	0.707	0.622	0.814	**0.836**
	Rec.	0.707	0.621	0.812	**0.832**
	F1	0.705	0.621	0.812	**0.830**
Weibo	Acc	0.736	0.794	0.873	**0.900**
	Prec.	0.608	0.802	0.824	**0.882**
	Rec.	0.588	0.791	0.815	**0.823**
	F1	0.595	0.791	0.820	**0.847**



Experiments

Research Questions

- Is the proposed model improving multimodal fake news detection by leveraging intra and inter-modality relationships?
- How effective are the extracted fragments and self-attention representations in improving the multimodal fake news detection?
- Can the proposed model identify the modality that aided in easy recognition of falsification in a particular news sample?

Experiments

Case Study

(a) False Context | (b) Fabricated Content | (c) False Connection | (d) Manipulated Content

Eiffel Tower lit up in Pakistan colours after yesterday's barbaric attacks in Lahore

Beware and Bewarned THe FIVe Headed Snake. A snake with Five heads

@Palestine_Pics: Syrian girl selling chewing gum in the streets of Jordan. | via @Trotsmoslim #FreeSyria

Ok... Who wants to tell President Bush that the library book he's reading is upside down?\n#bushlibrarybooks

0.7, 0.4

0.03, 0.8695

Conclusion
of LIIRM

- Presenting a novel framework that leverages intra and inter modality relationships for multimodal fake news detection.
- Proposed method comprises of two sub-modules.
  - Intra-modality feature extractor
    - BERT+Text-CNN & image fragments are obtained via bottom-up attention.
  - Inter-modality relationship extractor
    - Fuses multimodal features multiplicatively.

Comments

of LIIRM

- Datasets and comparison models are old version (classic).
  - MediaEval
  - EANN, MVAE, SpotFake (even no SpotFake+)
- Extract image fragments to obtain more useful informations.
  - May can utilized in my method.