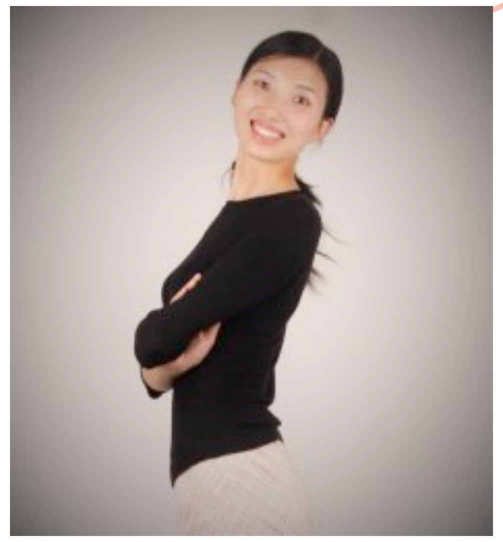
Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues



Peng Qi[1,2,3], Juan Cao[1,2], Xirong Li[4], Huan Liu[5], Qiang Sheng[1,2], Xiaoyue Mi[1,2], Qin He[6], Yongbiao Lv[6], Chenyang Guo[6], Yingchao Yu[6]

[1]Key Lab of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, China
[2]University of Chinese Academy of Sciences [3]Institute of Artificial Intelligence, Hebi, China
[4]Key Lab of DEKE, Renmin University of China, Beijing, China [5]Zhengzhou University, Zhengzhou, China
[6]Hangzhou ZhongkeRuijian Technology Co., Ltd., Hangzhou, China
{qipeng,caojuan,shengqiang18z,mixiaoyue19s}@ict.ac.cn,xirong@ruc.edu.cn,liuhuan_2012@hotmail.com,
{heqin,lvyongbiao,guochenyang,yuyingchao}@ruijianai.com

MM'21

220317 Chia-Chun Ho

ACM multimedia
Chengdu, China OCT 20-24 2021

Outline

Introduction

Fake News Detection

- The rising prevalence of fake news and its alarming real world impacts have motivated both academia and industry to develop automatic method to detect fake news.
- Traditional approaches typically focus on textual content.
- With the recent evolution of fake news from text-only posts to multimedia posts with photos and videos.
    - Approaches based on multimodal content demonstrate promising detection performance.
- In this paper, target on multimodal (text & image) fake news detection.

Introduction

Multimodal fake news detection

- Existing works model the multimodal content insufficiently.
- Most of them only preliminary model the basic semantics of the images as a complement of the text.
  - Ignoring the characteristics of multimodal fake news.
  - Some prior works obtain the multimodal representations by simply concatenating the textual features with visual features extracted from pre-trained model.
- To make up for this omission, explore 3 valuable text-image correlations in multimodal fake news, which provide diverse multimodal clues.

Introduction

Text & images have inconsistent entities

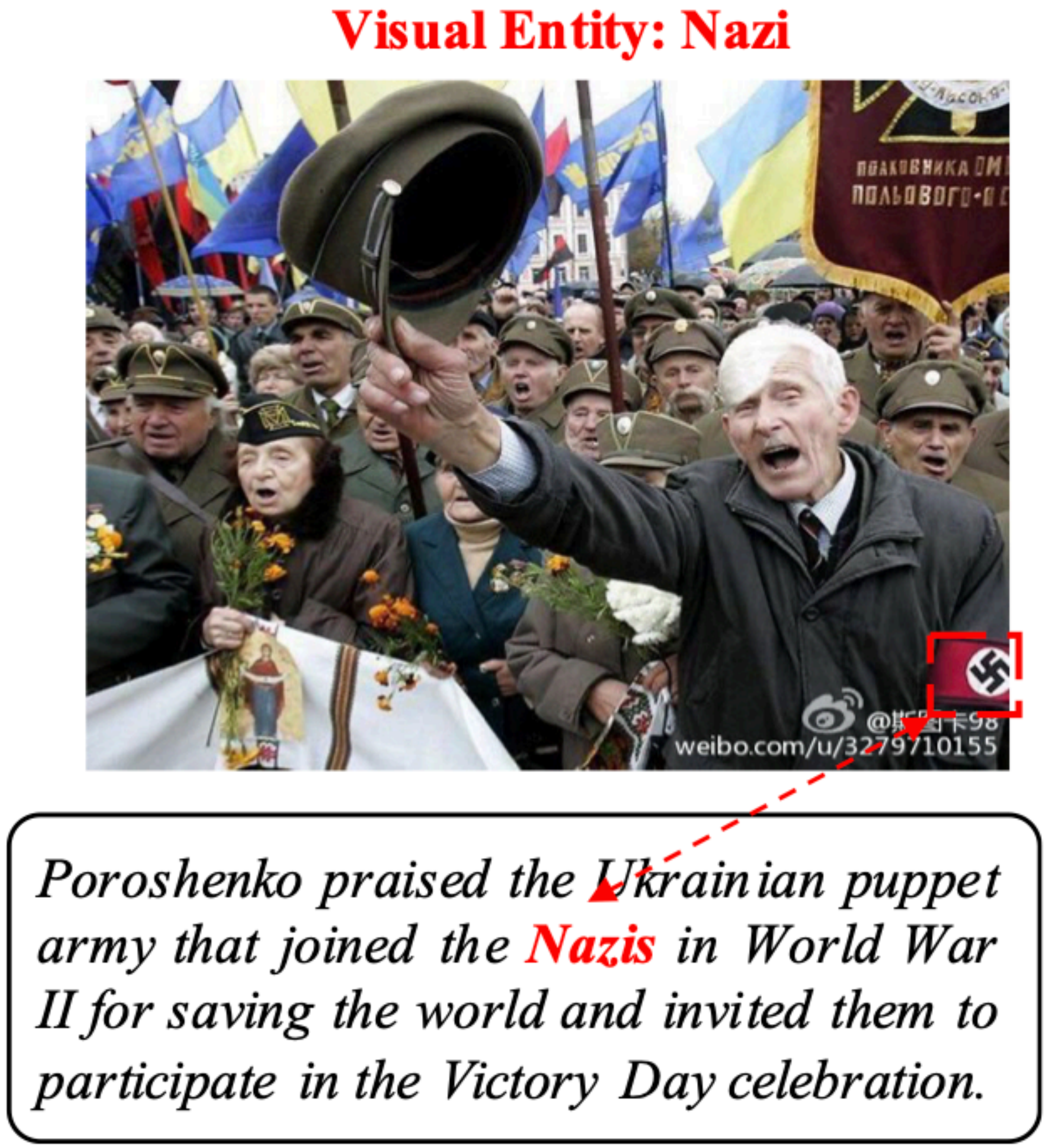
- It's a potential indicator for multimodal fake news.
- Wrongly reposting outdated images is a typical way to make up fake news.
- It's difficult to find both semantically pertinent and non-manipulated images to support these non-factual stories.
- As shown in figure, the text describes a piece of news about "Dallas Jones" while the attached image is the arrest scene of another person.

(a) Entity Inconsistenty

Introduction

Text & images enhance each other by spotting the important features

- News text and images are related in high-level semantics, and the aligned parts usually reflect the key elements of news.
- In this kind of news, text provides main clues for detection, while image select the key clues in the text.
- As figure shows, the Nazi flag in the image corresponds to the important entity “Nazi” in the text, which is the key controversial point of this news post.



Introduction

Embedded text in images provides complementary information for original text

- According to statistics on the Weibo dataset, more than 20% of multimodal fake news spreads in the form of image.
- This refers to news that the embedded text in the image tells the complete fake news story while the original text often is comment.
- In this kind of fake news, the clues lie in the combination of the original text and the embed text in the image.

(c) Text Complementation

Introduction

Another challenge of fusing multimodal information

- Lies in the heterogeneity of multimodal data.
- Current works focus on the general objects of news images by pre-trained model.
  - News text is in a more abstract semantic level based on named entities.
  - Due to this semantic gap, current works are hard to reason effectively between text and images for exploring multimodal clues.
- For example on previous figure, can't reveal the inconsistency as clues to detect this news as fake if only recognize the celebrity in the images as "person" instead of "Cuba Gooding Jr."

Introduction

Another challenge of fusing multimodal information

- Import the visual entities to model the high-level semantics of news images.
- Visual entities consist of words describing named entities recognized from the images (celebrity & landmark) and some news-related visual concepts.
- They are important for mining the clues because they
  - Contain rich visual semantics and thus help understand the multimodal news.
  - Bridge the high-level semantic correlations of news text and images.

Introduction
EM-FEND

- Propose a framework of multimodal fake news detection, named as EM-FEND.
    - Entity-enhanced Multimodal FakE News Detection
- Fuses diverse multimodal clues to detect multimodal fake news.
- In feature extraction, in addition to extract the basic visual features through fine-tuned VGG19, explicitly extract visual entities and the embedded text in images to modal the high-level visual semantics.
- Besides, explicitly extract textual entities to capture the key elements of the news events.

Introduction

EM-FEND

- In the stage of fusion, model 3 types of cross-modal correlations in multimodal fake news to fuse diverse multimodal clues for detection.
    - Text Complementation: concatenate the original text and the OCR text in images as the composed text and feed it into BERT to obtain the fused textual features.
    - Mutual Enhancement: use co-attention transformers between text with visual and visual CNN features.
    - Entity Inconsistency: by calculating the similarity of textual and visual entities.
- Then fuse the above multimodal features by concatenation to feed for classification.

Introduction
Contributions

- Find 3 valuable text-image correlations in multimodal fake news, and propose a unified framework to fuse these multimodal clues simultaneously.
- To authors' best knowledge, EM-FEND is the first import the visual entities into multimodal fake news detection.
    - Helps to understand the news-related high-level semantics of images and bridge the high-level semantic correlations of news text and images.
- Both offline and online evaluations demonstrate the superiority of proposed model compared to the SOTAs.

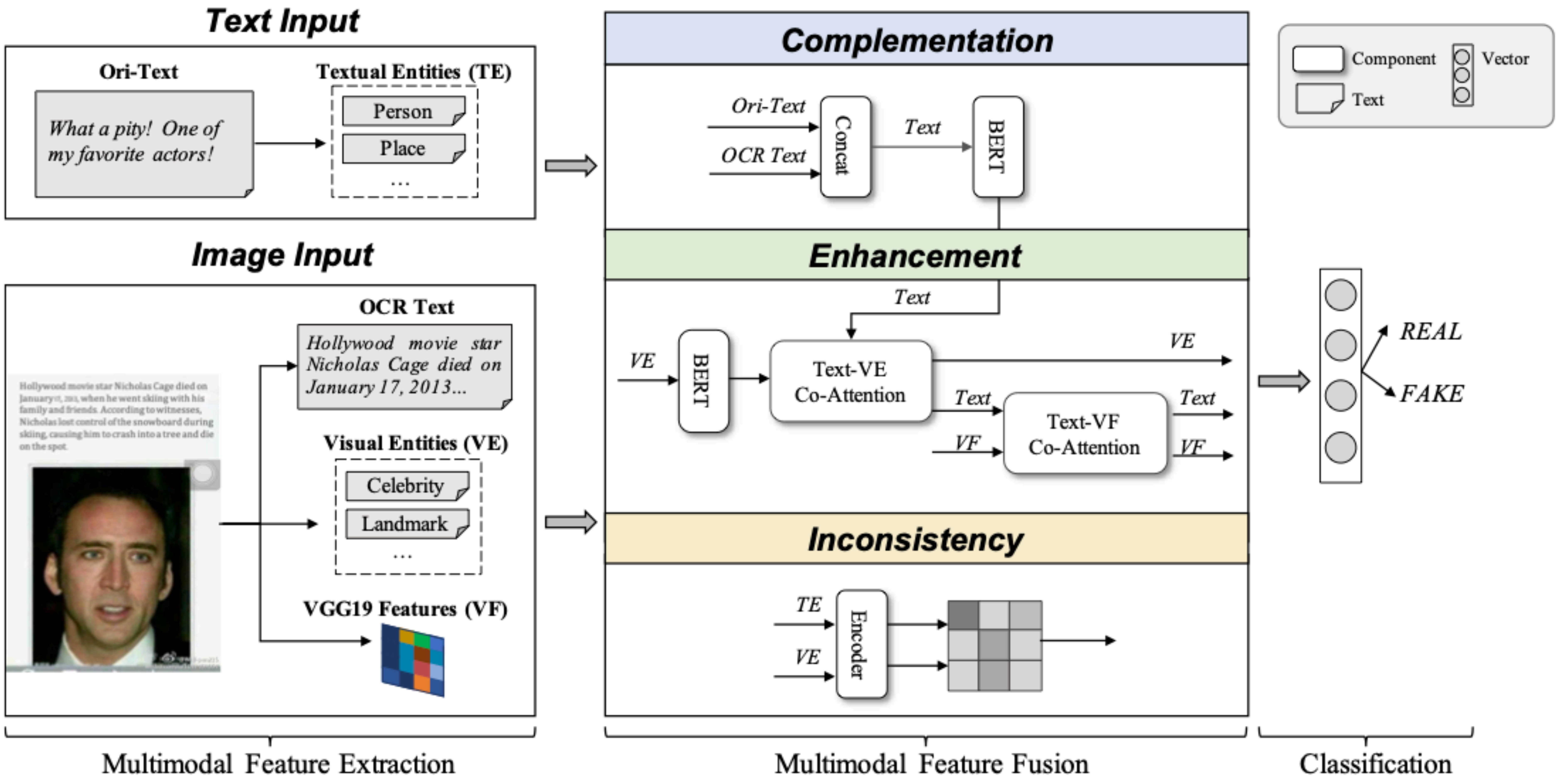
Related Works

Fake News Detection

Methods	Backbone			Cross-modal Correlations		
	Text	Image	Fusion	*inconsistency*	*enhancement*	*text complementation*
EANN[26]	Text-CNN	VGG19	concat	-	-	-
metaFEND[27]	Text-CNN	VGG19	concat	-	-	-
MVAE[7]	Bi-LSTM	VGG19	variational autoencoder	-	-	-
SpotFake[23]	BERT	VGG19	concat	-	-	-
SAFE[34]	Text-CNN	image2sentence +Text-CNN	concat+multi-loss	text-imagecaption	-	-
MCNN[29]	BERT +Bi-GRU	ResNet50 +Attention	attention+multi-loss	text-visfea	-	-
attRNN[9]	Bi-LSTM	VGG19	neuron-level attention	-	text->visfea	-
MKEMN[32]	Bi-GRU	VGG19	attention +multi-channel CNN	-	text->visfea	-
CARMN[24]	BERT	VGG19	co-attention transformer +multi-channel CNN	-	text<->visfea	-
KMGCN[28]	-	YOLOv3	GCN	-	text<->objects	-
EMAF[12]	BERT	Faster-RCNN	Capsule	-	text<->object fea	-
EM-FEND(ours)	BERT	VGG19 +entity detector +OCR model	co-attention transformer	text-visentity	text<->visfea text<->visentity	+

Methodology

EM-FEND



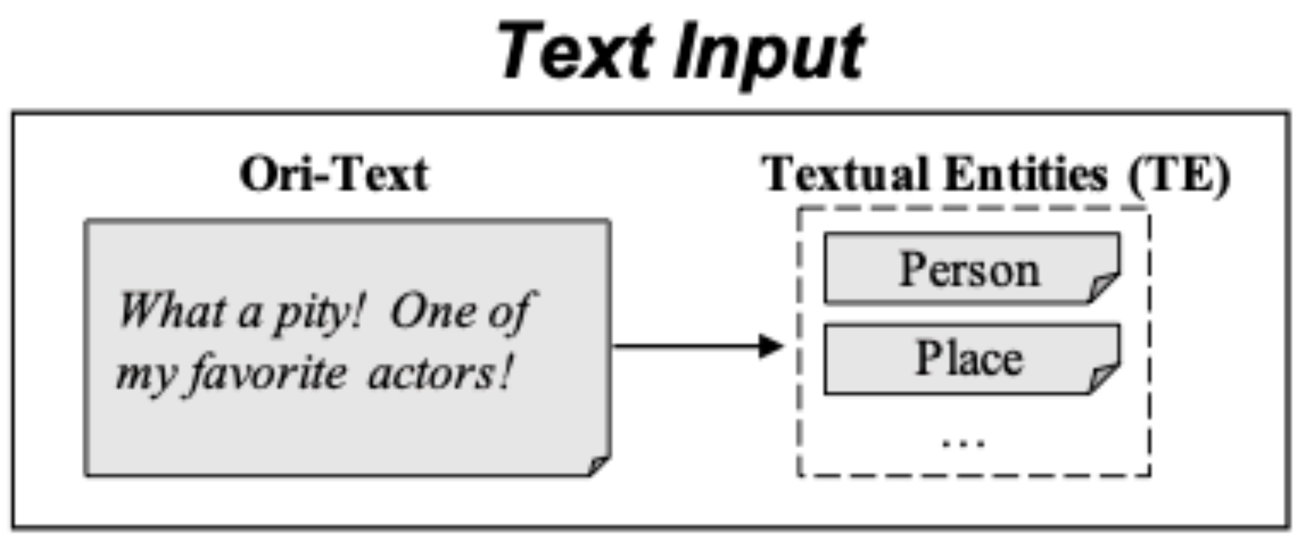
Methodology

EM-FEND

- EM-FEND includes 3 modules to fuse diverse multimodal clues for FND.
  - Multimodal feature extraction
    - Extract textual & visual entities, embedded text in the image, visual CNN features.
  - Multimodal feature fusion
    - Correlations: entity inconsistency, mutual enhancement, & complementation
  - Classification
    - Use the obtained multimodal representation to perform binary classification.

Methodology

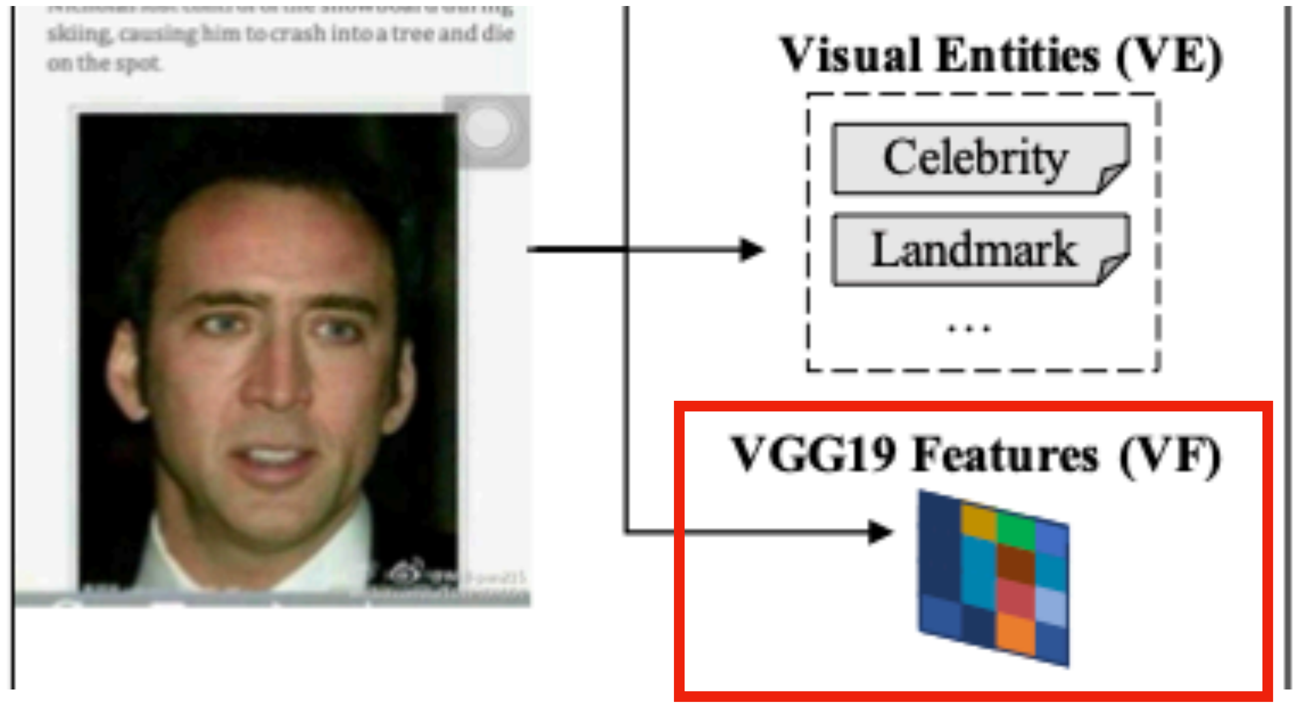
Feature Extraction: Text Entities



- As a special narrative style, news usually contains named entities such as persons and locations.
- These entities are of importance in understanding the news semantics and also helpful in detecting fake news.
- Explicitly extract the person entities P_T and location entities L_T by recognizing corresponding proper nouns in the text.
- For better understanding the news event, employ part-of-speech (POS) tagging to extract all nouns as a general textual context C_T.

Methodology

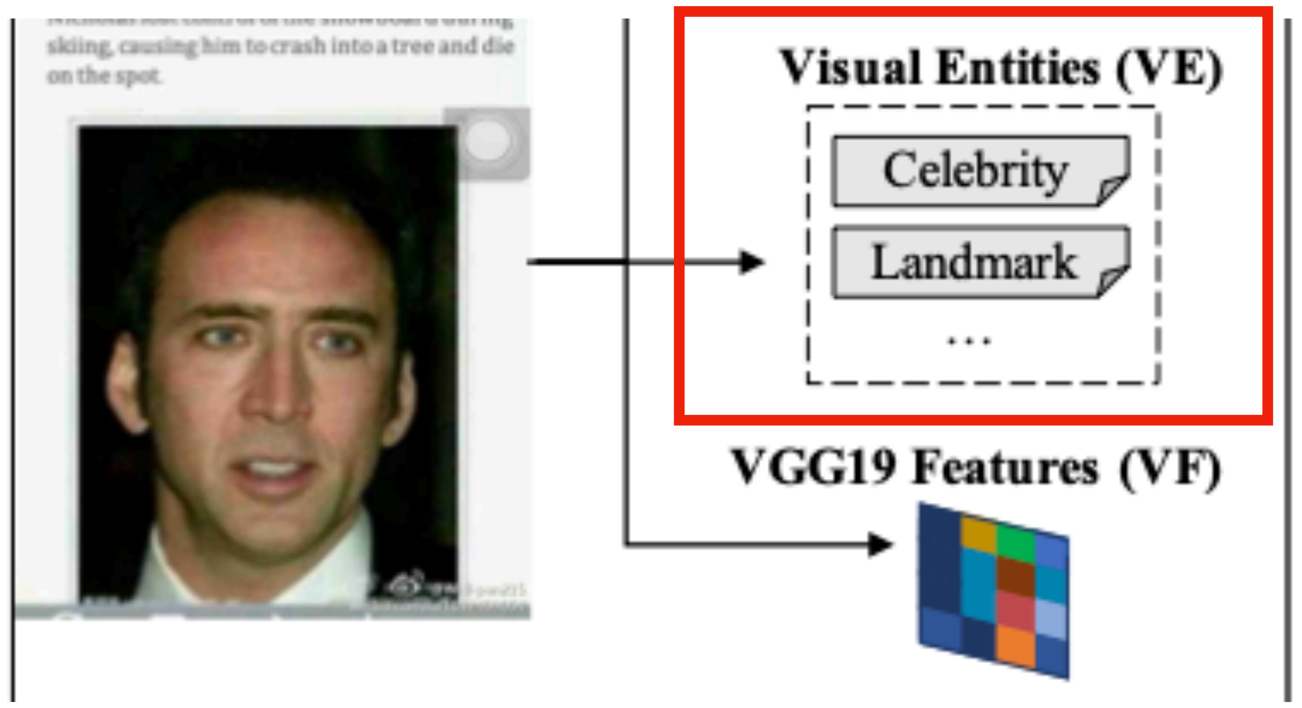
Feature Extraction: Visual CNN Features



- Follow previous works, adopt VGG19 to extract the visual features.
- The difference is fine-tune the pre-trained VGG19 on the given dataset to flexibly capture the low-level characteristics of the images from the specific data source to help detection.
- Considering that different regions in the image may show different patterns.
    - Split the original image into 7×7 regions, and then obtain the corresponding visual features sequence $H_V = [r_1, \cdots, r_n], n = 49$.

Methodology

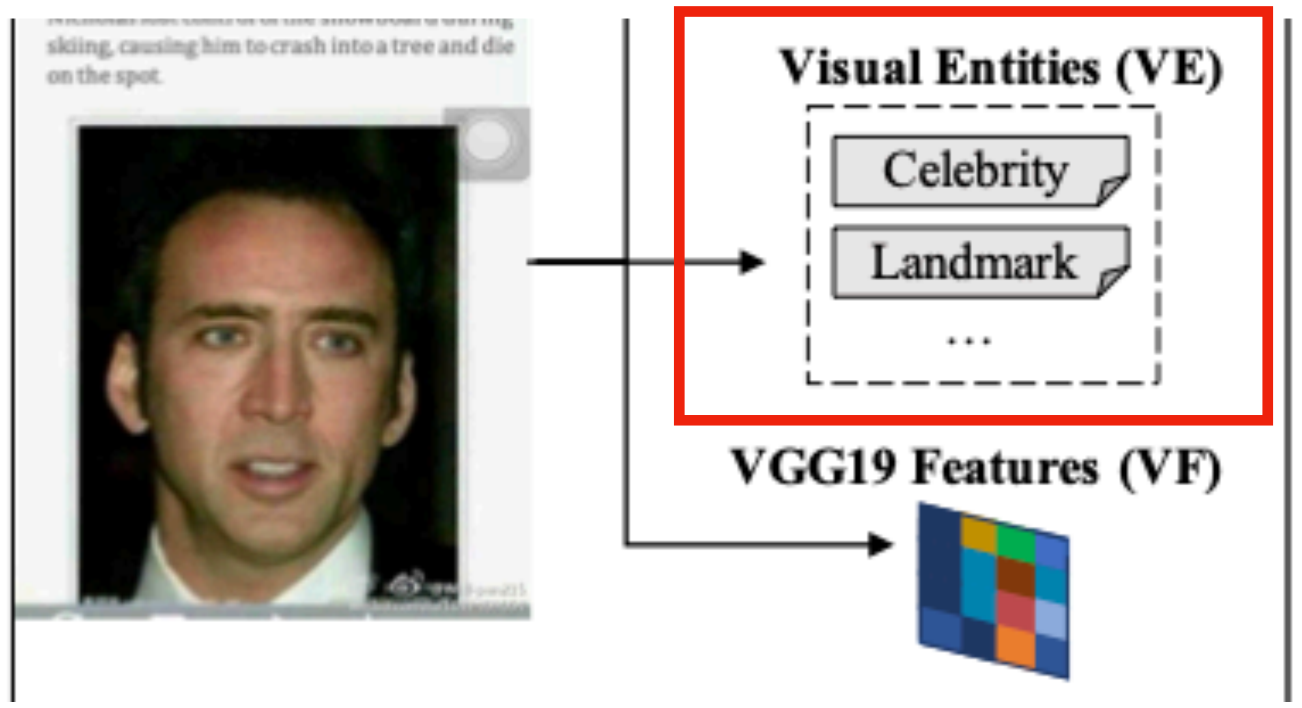
Feature Extraction: Visual Entities



- Similar to text, news image also contain newsworthy visual entities.
- Specifically, extract 4 types of visual entries:
  - Celebrities & landmarks
  - Organization (e.g. Nazi, Buddhism and police, by detecting flags of clothes)
  - Eye-striking visual concepts (e.g. violence, bloodiness, and disaster)
  - General objects and scenes

Methodology

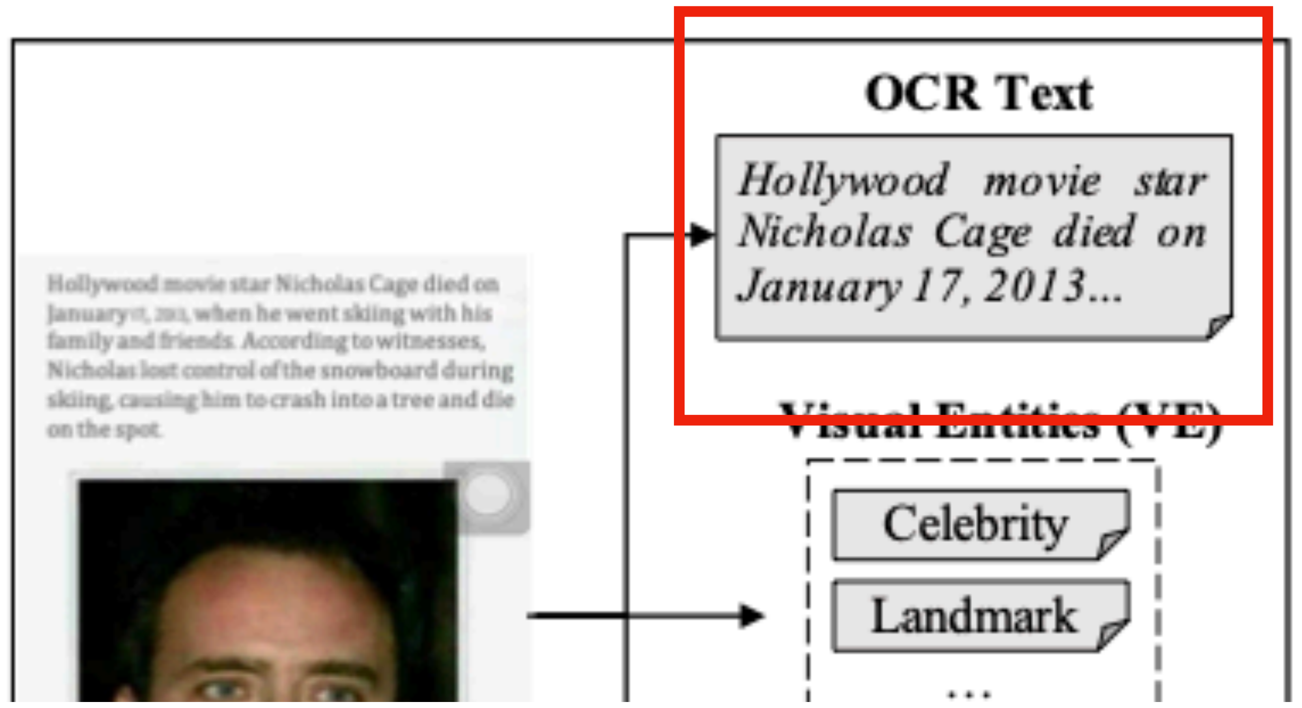
Feature Extraction: Visual Entities



- Due to the high accuracy requirements for pretrained models and the lacking of relevant publicity available datasets.
- Use public APIs to detect visual entities instead of re-implement these models.
- Finally, obtain the
  - person entities P_V
  - location entities L_V
  - other news related visual concept as general image context C_V

Methodology

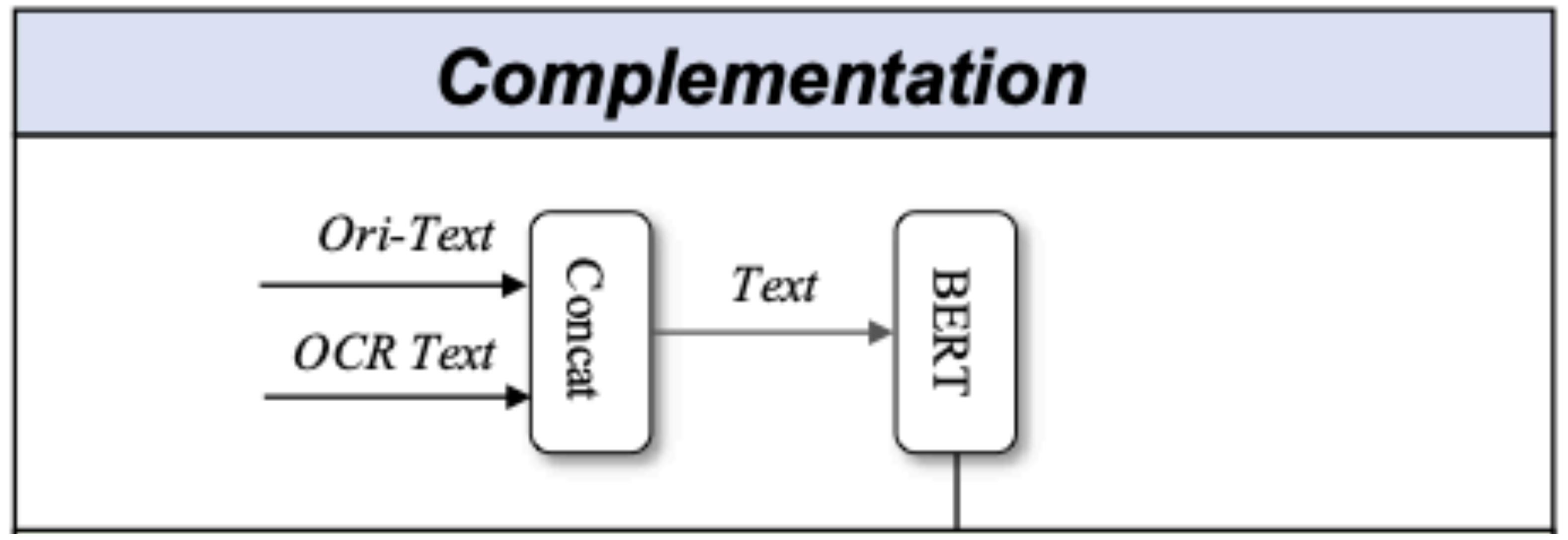
Feature Extraction: Embedded Text



- In addition to the original input text, text embedded in images is also important.
    - It usually contains important information missed by the original text.
- Extract the embedded text O of the input image by applying the OCR model.

Methodology

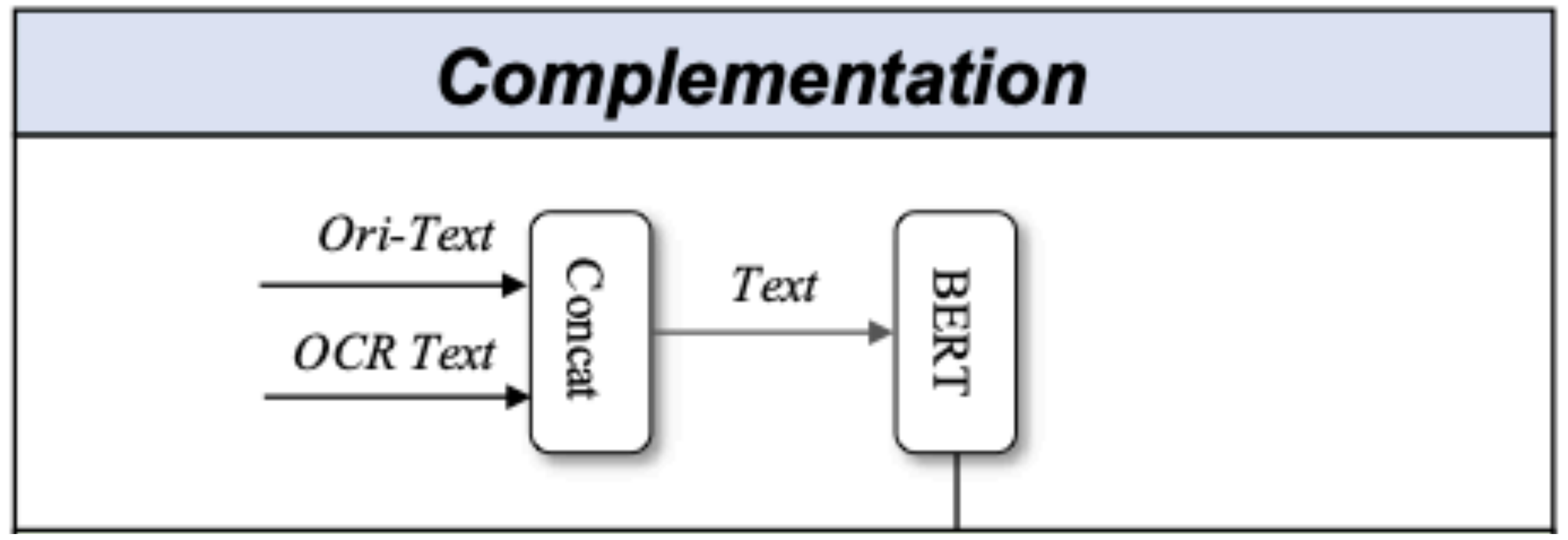
Feature Fusion: Text Complementation



- As the main body of multimodal news, text provides rich clues for the judgement of news credibility.
- For fake news in social media, in addition to the original text, the embedded text in images is also important in understanding the news semantics and providing clues for detection.
- In many situations, the key clues for detection lie in the embedded text, while the original text is just a comment about the news event.
- Therefore, the original and the embedded text should be modeled jointly to obtain the whole semantics of the news events.

Methodology

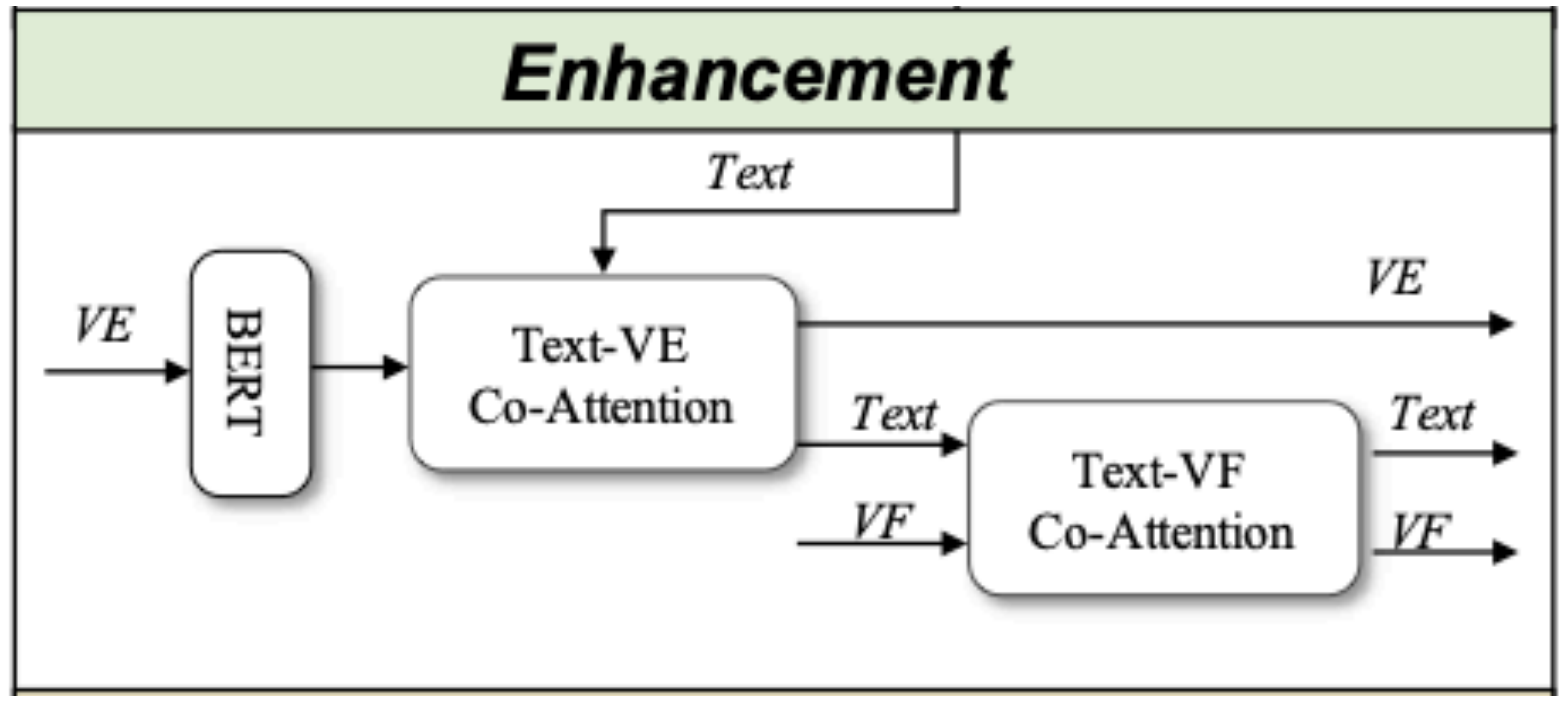
Feature Fusion: Text Complementation



- Most existing methods use recurrent or CNNs to model the contextual information of the text sequence.
- Recently, pre-trained LM have shown strong ability in modeling text.
- Thus, feed the original text T and embedded text O into the pre-trained BERT.
    - $H_T = \text{BERT}([CLS]T[SEP]O[SEP])$
- Then obtain the textual feature $H_T = [w_1, \cdots, w_n]$

Methodology

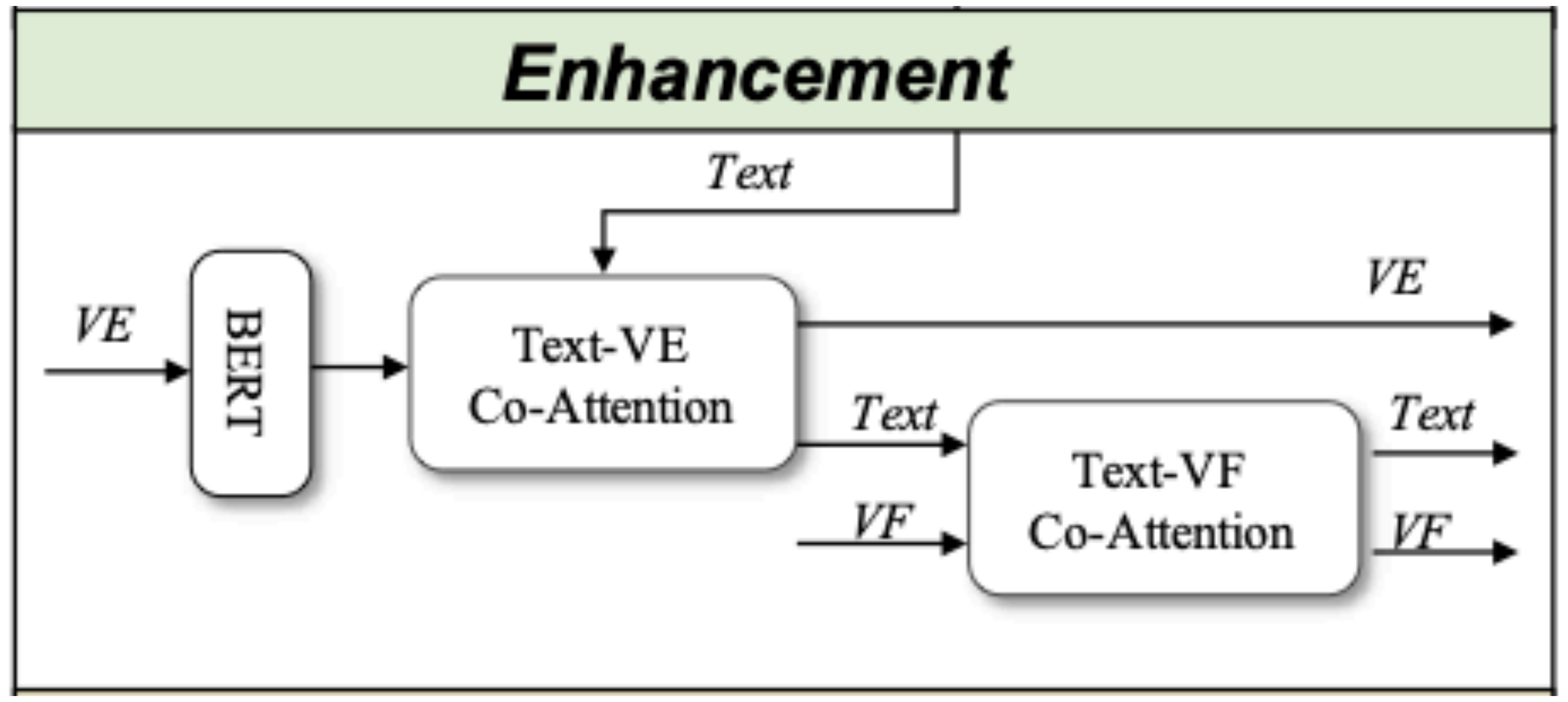
Feature Fusion: Mutual Enhancement



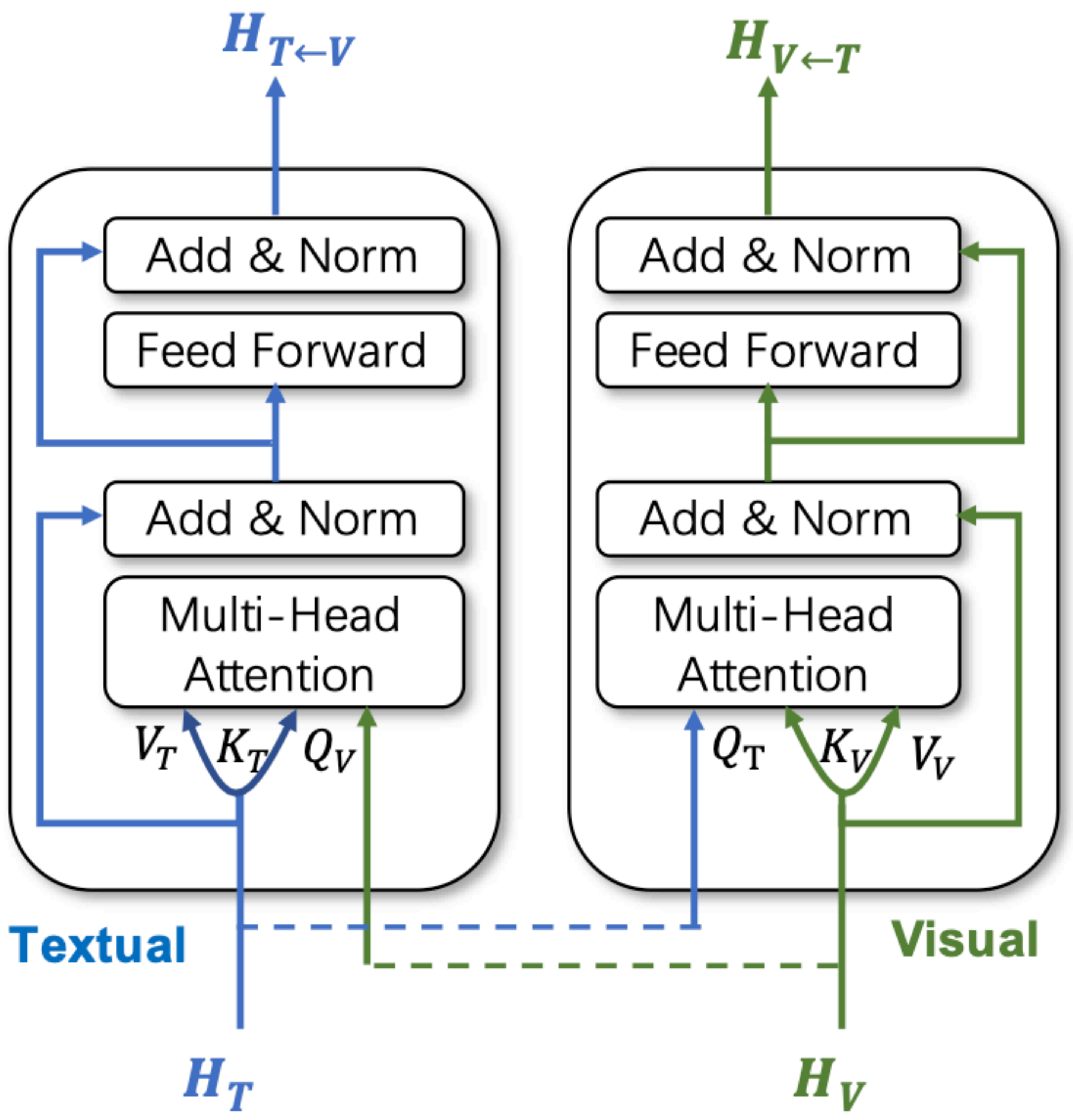
- Important news elements mentioned in the text are usually illustrated and emphasized by images and vice versa.
- Thus, the text and images could spot the important features respectively by aligning with each other.
- Inspired by the success of the co-attention mechanism in VQA task.
  - Use the multimodal co-attention transformer between textual & visual entities & visual CNN features to model multimodal alignment at different visual levels.

Methodology

Multimodal Co-attention Transformer (MCT)

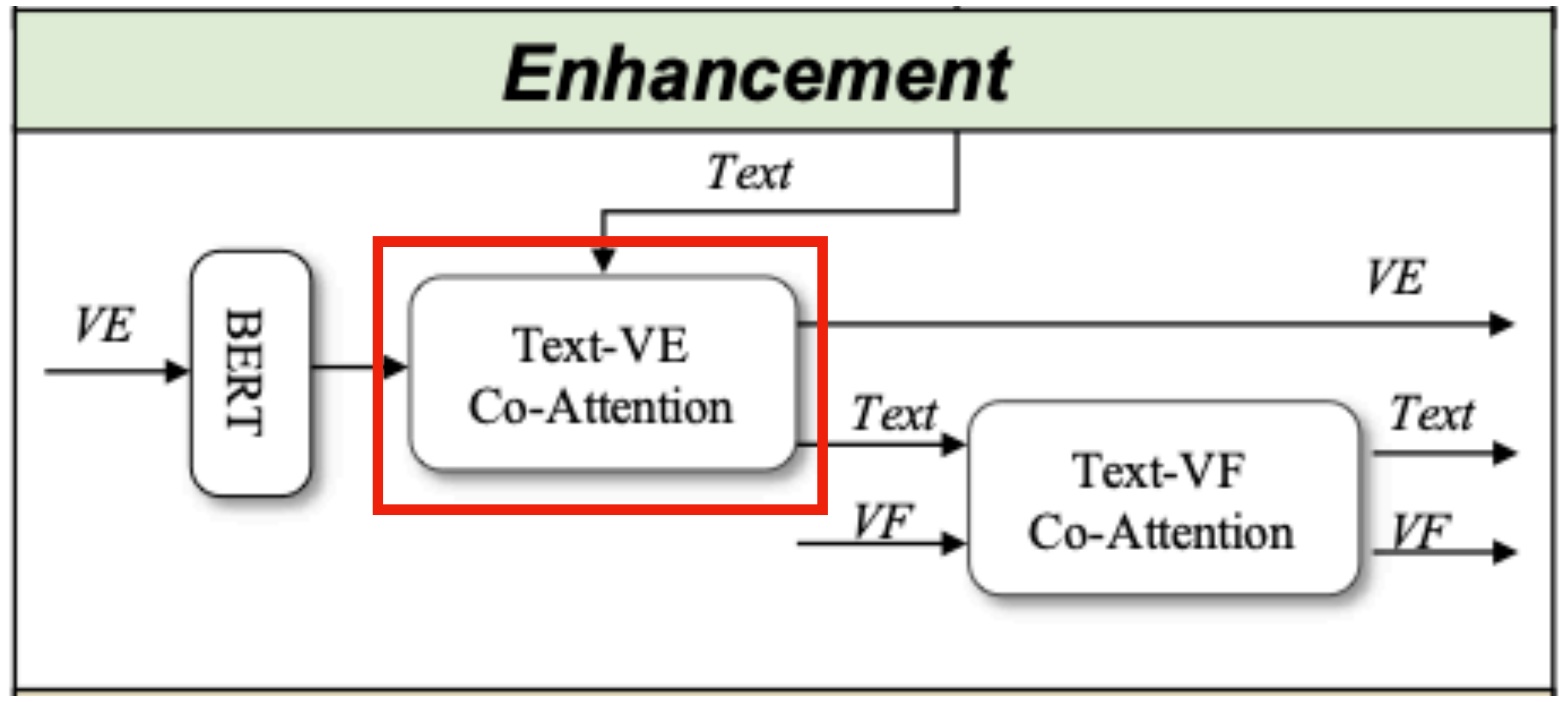


- Use a two-stream transformer to process the textual & visual information simultaneously.
- Modify the standard query-conditioned key-value attention mechanism to develop a multimodal co-attentional transformer module.
- The queries from each modality are passed to the other modality's multi-headed attention block.
- Consequentially this transformer produces image-enhanced textual features and text-enhanced visual features.



Methodology

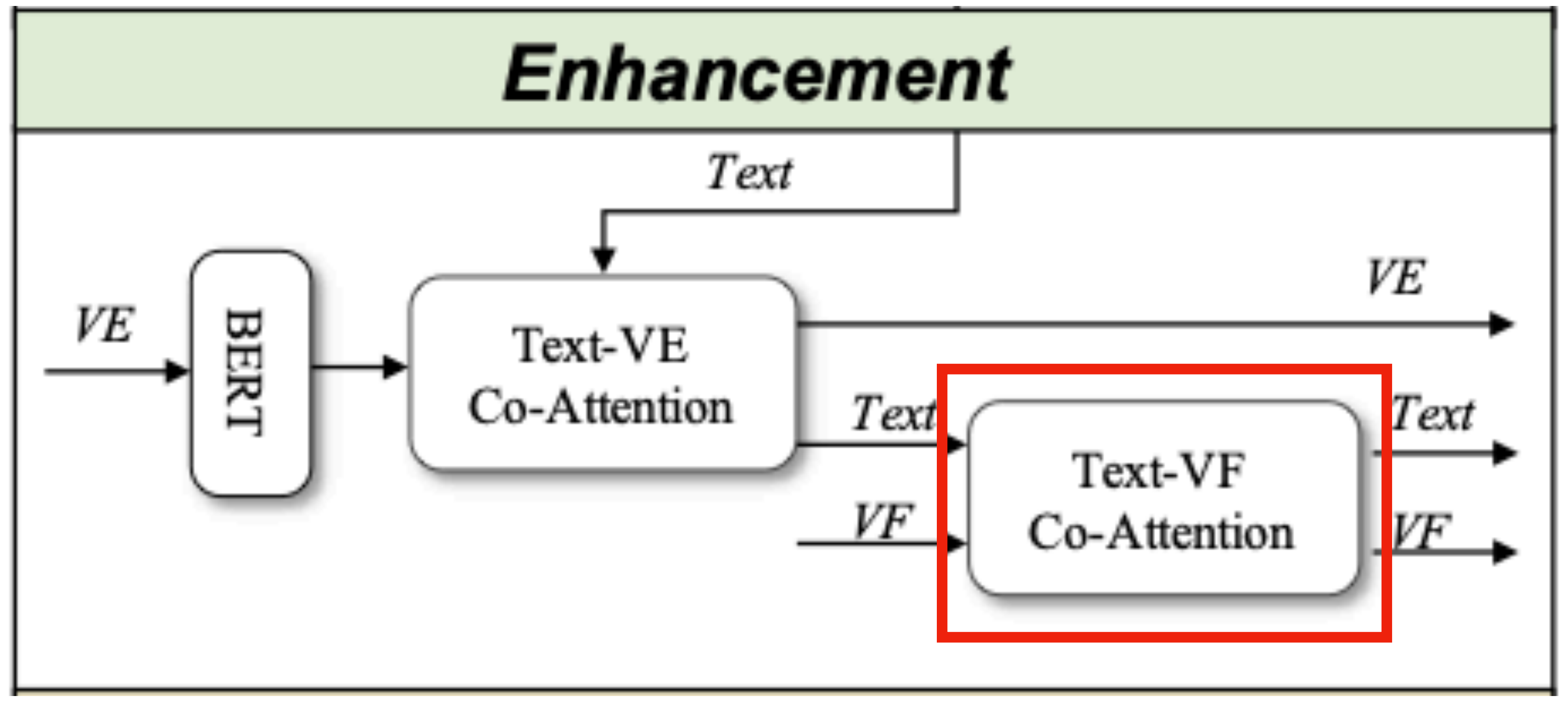
MCT between Textual & Visual Entities



- After obtaining the visual entities VE, employ the pre-trained BERT to obtain their embeddings H_{VE}.
- Thus the textual and visual entities' embeddings could be fused in similar BERT-constructed feature spaces, alleviating the problem of multimodal feature heterogeneity.
- The aligned words and visual entities usually reflect the key elements of the news, thus use MCT to fuse these features. Feed the H_T & H_{VE} into first MCT to obtain the textual representation enhanced by visual entities $H_{T \leftarrow VE}$ and vice versa $H_{VE \leftarrow T}$.
- Apply the average operation to obtain the final representation of visual entities x_{ve}.

Methodology

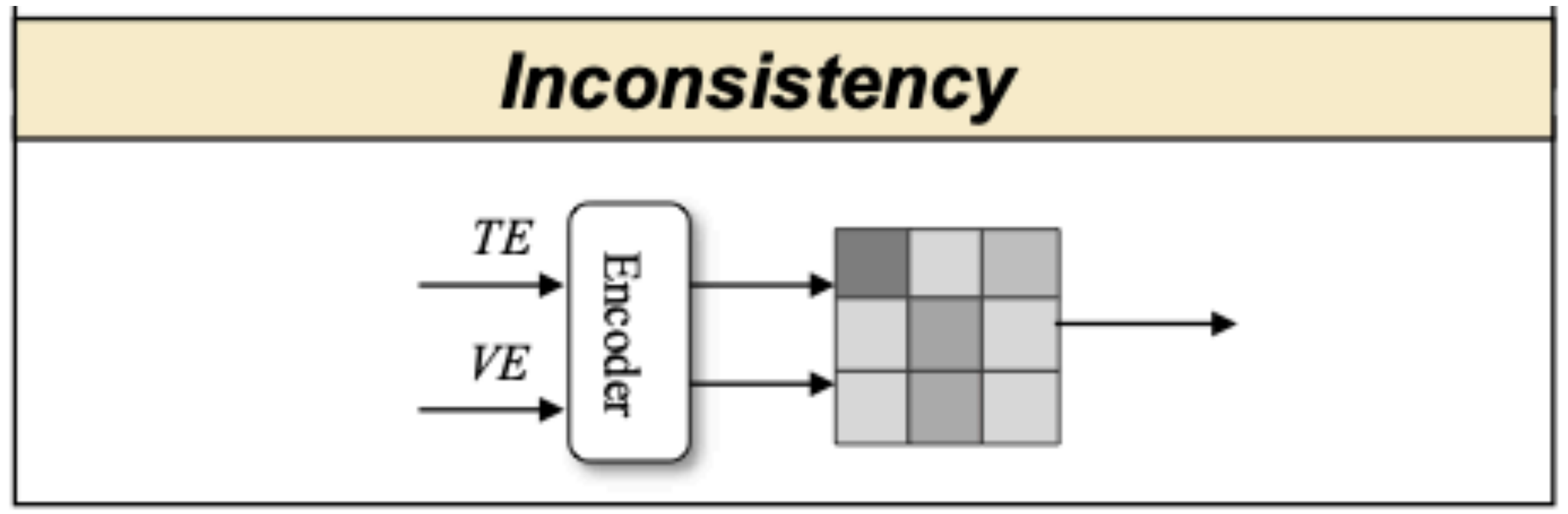
MCT between Textual & Visual CNN Features



- Visual entities focus on the local high-level semantics of the images.
  - Ignoring the global low-level visual features.
- As a supplement, use MCT to model the correlations between textual & visual CNN features.
- Feed the $H_{T \leftarrow VE}$ & H_V into second MCT to obtain the textual representation enhanced by both visual entities and visual CNN features $H_{T \leftarrow (VE,V)}$ and vice versa $H_{V \leftarrow T}$.
- Apply the average operation to obtain the final representation of the text x_t and image x_v.

Methodology

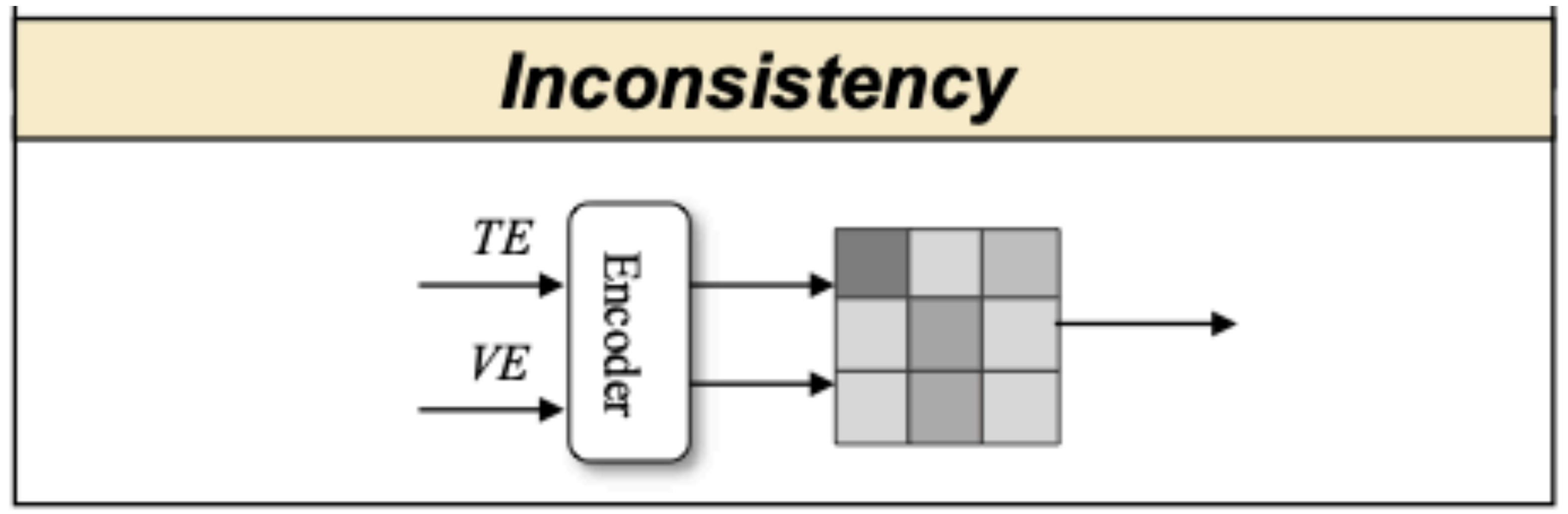
Entity Inconsistency Measurement



- Measure the multimodal entity inconsistency of person, location, and a more general event context.
- There are 2 challenges for this measurement:
  - The heterogeneity of textual and visual features.
    - Calculate similarity on textual feature space based on their embeddings.
  - News text usually contains more entities and information than the companying images, and thus some textual entities could be without the aligned visual entities.
    - Consider entity inconsistent only when there are no aligned multimodal entities.

Methodology

Entity Inconsistency Measurement



- Taking person entity as an example, define the cross-modal person similarity as the maximum similarity among all pairs of textual and visual person entities.
- Since neural network have inevitable errors when detecting visual entities, the confidence is considered when computing the similarity.
- Calculate the cross-modal person similarity as
- $$x_s^p = \max_{t \in T_p}\left(\sum_{v \in V_p} \rho(\boldsymbol{v}) \frac{\boldsymbol{t} \cdot \boldsymbol{v}}{\|\boldsymbol{t}\| \, \|\boldsymbol{v}\|}\right)$$
- Similarly, compute the x_s^l, x_s^c then concatenate them to form the entity consistency x_s.

Methodology
Classification

- Finally, concatenate the final representation of the text x_t, visual entities x_{ve}, image x_v, and the multimodal entity consistency feature x_s to obtain final representation x_m.
  - $x_m = \text{concat}(x_t, x_{ve}, x_v, x_s)$
- Use a fully connected layer with softmax activation to project x_m into the target space.
  - $p = \text{softmax}(Wx_m + b), p = [p_0, p_1]$ [real, fake]
- Use binary cross-entropy to minimize the loss.
  - $\mathscr{L}_p = -[y \log p_0 + (1 - y) \log p_1]$

Experiments
Datasets

- Chinese: Weibo-16
  - 4749 fake : 4779 real
- English: Long news article on news websites
  - 2844 fake : 2825 real
- Use K-means to find the common events and split data into training, validation, testing set based on event clusters to ensure that there is no overlap among these sets.
- Training 3: Validation 1: Testing 1

Experiments

Baselines

- Single-modality Methods: Bi-LSTM, BERT, VGG19
- Multimodal Methods:
  - att-RNN: use RNN with attention mechanism to fuse text and visual information.
  - MVAE: utilizes a multimodal variational autoencoder trained jointly detector to learn representation.
  - MKN: retrieve concepts of textual entities from external knowledge graphs.
  - SAFE: translate image into sentence and compute the relevance based on sentence similarity.
  - SpotFake: concat the BERT textual and VGG19 visual feature for classification.
  - CARMN: propose a cross-modal attentions residual network to fuse multimodal features.

Experiments

Baselines

- Considering that using pre-trained LM to extract textual features usually improves the detection performance of models event without significant changes on the model structure.

- Design a reduced variant of the proposed EM-FEND model to ensure the fairness of comparisons.

    - EM-FEND-base: use Bi-LTSM with pre-trained word2vec to replace BERT in EM-FEND.

Experiments

Evaluation Questions

- EQ1: Can EM-FEND improve the classification performance of distinguishing multimodal fake and real news?
- EQ2: How effective are various visual features (especially visual entities) and cross-modal correlations in improving the performance of EM-FEND?
- EQ3: How does EM-FEND perform in online fake news detection?

Experiments

Evaluation Questions

- EQ1: Can EM-FEND improve the classification performance of distinguishing multimodal fake and real news?
- EQ2: How effective are various visual features (especially visual entities) and cross-modal correlations in improving the performance of EM-FEND?
- EQ3: How does EM-FEND perform in online fake news detection?

Experiments

Performance Comparison

- EM-FEND is much better than other methods on both datasets.
- It's validates that EM-FEND can effectively capture important multimodal clues.

	Methods	Acc.	Prec.	Recall	F1
Chinese	Bi-LSTM	0.785	0.851	0.692	0.763
	BERT	0.830	**0.977**	0.675	0.798
	VGG19	0.730	0.789	0.626	0.698
	attRNN-[9]	0.808	0.882	0.711	0.787
	MVAE[7]	0.797	0.827	0.751	0.787
	MKN[32]	0.805	0.865	0.722	0.787
	SAFE[34]	0.790	0.886	0.665	0.760
	EM-FEND-base (Ours)	0.852	0.841	0.853	0.847
	SpotFake[23]	0.852	0.854	0.850	0.852
	CARMN [24]	0.865	0.933	0.774	0.846
	EM-FEND (Ours)	**0.904**	0.897	**0.904**	**0.901**
English	Bi-LSTM	0.864	0.877	0.843	0.859
	BERT	0.873	0.869	0.875	0.872
	VGG19	0.773	0.783	0.747	0.764
	attRNN-[9]	0.872	0.861	0.882	0.871
	MVAE[7]	0.879	0.902	0.848	0.874
	MKN[32]	0.889	0.846	0.929	0.886
	SAFE[34]	0.909	0.922	0.890	0.906
	EM-FEND-base (Ours)	0.943	0.926	0.961	0.943
	SpotFake[23]	0.899	0.879	0.923	0.901
	CARMN [24]	0.937	0.934	0.940	0.937
	EM-FEND (Ours)	**0.975**	**0.978**	**0.973**	**0.975**

Experiments

Performance Comparison

- Methods based on textual modality are better than the visual modality.
  - Proving that the textual modality more rich clues than images.
- Then multimodal methods are generally better than single-modality ones.
  - Indicating the complementary of multimodal features.

	Methods	Acc.	Prec.	Recall	F1
Chinese	Bi-LSTM	0.785	0.851	0.692	0.763
	BERT	0.830	**0.977**	0.675	0.798
	VGG19	0.730	0.789	0.626	0.698
	attRNN-[9]	0.808	0.882	0.711	0.787
	MVAE[7]	0.797	0.827	0.751	0.787
	MKN[32]	0.805	0.865	0.722	0.787
	SAFE[34]	0.790	0.886	0.665	0.760
	EM-FEND-base (Ours)	0.852	0.841	0.853	0.847
	SpotFake[23]	0.852	0.854	0.850	0.852
	CARMN [24]	0.865	0.933	0.774	0.846
	EM-FEND (Ours)	**0.904**	0.897	**0.904**	**0.901**
English	Bi-LSTM	0.864	0.877	0.843	0.859
	BERT	0.873	0.869	0.875	0.872
	VGG19	0.773	0.783	0.747	0.764
	attRNN-[9]	0.872	0.861	0.882	0.871
	MVAE[7]	0.879	0.902	0.848	0.874
	MKN[32]	0.889	0.846	0.929	0.886
	SAFE[34]	0.909	0.922	0.890	0.906
	EM-FEND-base (Ours)	0.943	0.926	0.961	0.943
	SpotFake[23]	0.899	0.879	0.923	0.901
	CARMN [24]	0.937	0.934	0.940	0.937
	EM-FEND (Ours)	**0.975**	**0.978**	**0.973**	**0.975**

Experiments

Performance Comparison

- Pre-trained LM (e.g. BERT) can improve the performance of proposed method.
    - Due to the strong ability of transformers in modeling context and the abundant knowledge injected in the pre-trained models.

	Methods	Acc.	Prec.	Recall	F1
Chinese	Bi-LSTM	0.785	0.851	0.692	0.763
	BERT	0.830	**0.977**	0.675	0.798
	VGG19	0.730	0.789	0.626	0.698
	attRNN-[9]	0.808	0.882	0.711	0.787
	MVAE[7]	0.797	0.827	0.751	0.787
	MKN[32]	0.805	0.865	0.722	0.787
	SAFE[34]	0.790	0.886	0.665	0.760
	EM-FEND-base (Ours)	0.852	0.841	<u>0.853</u>	0.847
	SpotFake[23]	0.852	0.854	0.850	<u>0.852</u>
	CARMN [24]	<u>0.865</u>	<u>0.933</u>	0.774	0.846
	EM-FEND (Ours)	**0.904**	0.897	**0.904**	**0.901**
English	Bi-LSTM	0.864	0.877	0.843	0.859
	BERT	0.873	0.869	0.875	0.872
	VGG19	0.773	0.783	0.747	0.764
	attRNN-[9]	0.872	0.861	0.882	0.871
	MVAE[7]	0.879	0.902	0.848	0.874
	MKN[32]	0.889	0.846	0.929	0.886
	SAFE[34]	0.909	0.922	0.890	0.906
	EM-FEND-base (Ours)	<u>0.943</u>	0.926	<u>0.961</u>	<u>0.943</u>
	SpotFake[23]	0.899	0.879	0.923	0.901
	CARMN [24]	0.937	<u>0.934</u>	0.940	0.937
	EM-FEND (Ours)	**0.975**	**0.978**	**0.973**	**0.975**

Experiments

Evaluation Questions

- EQ1: Can EM-FEND improve the classification performance of distinguishing multimodal fake and real news?
- EQ2: How effective are various visual features (especially visual entities) and cross-modal correlations in improving the performance of EM-FEND?
- EQ3: How does EM-FEND perform in online fake news detection?

Experiments
Ablation Study

- Design several internal models for comparison, which is simplified variations of EM-FEND with certain visual features removed:
  - w/o visual entities: w/o visual entities extraction, and the following MCT with textual feature and entity inconsistency measurement module.
  - w/o OCR text
  - w/o fine-tune VGG feature: replace by pre-trained VGG19 w/o fine-tuning.

Experiments

Ablation Study

	Methods	Acc.	Prec.	Recall	F1
Chinese	EM-FEND	**0.904**	0.897	**0.904**	**0.901**
	w/o visual entities	0.886	**0.930**	0.823	0.873
	w/o OCR text	0.882	0.902	0.845	0.873
	w/o FT VGG feature	0.773	0.783	0.747	0.764
English	EM-FEND	**0.975**	**0.978**	**0.973**	**0.975**
	w/o visual entities	0.953	0.954	0.950	0.952
	w/o OCR text	0.970	0.967	0.972	0.969
	w/o FT VGG feature	0.970	0.954	0.988	0.971

- Most important features are different: VGG in Chinese, visual entities in English.
- Result from the differences in source between these two datasets.
- Chinese dataset more likely to show low image quality by wide propagation.
  - Low-level visual features → VGG-19
- English dataset from the formal news website, has high-quality and informative image.
  - High-level visual features → visual entities
- Proves the generalization ability of EM-FEND in detecting different types of fake news.

Experiments
Ablation Study

- Similarly, design the following variants of EM-FEND to prove the effectiveness of different cross-modal correlations:
    - w/o co-attention-ve: w/o MCT between textual & visual entities.
    - w/o co-attention-vf: w/o MCT between textual & visual CNN features.
    - w/o entity inconsistency measurement

Experiments

Ablation Study

	Methods	Acc.	Prec.	Recall	F1
Chinese	EM-FEND	**0.904**	0.897	**0.904**	**0.901**
	w/o entity consistency	0.899	**0.932**	0.849	0.889
	w/o co-attention-ve	0.890	0.914	0.851	0.881
	w/o co-attention-vf	0.886	0.901	0.855	0.878
English	EM-FEND	**0.975**	**0.978**	**0.973**	**0.975**
	w/o entity consistency	0.962	0.977	0.945	0.961
	w/o co-attention-ve	0.959	0.953	0.966	0.959
	w/o co-attention-vf	0.930	0.937	0.920	0.928

- The accuracy is lower than the complete model by at least 1.4% in accuracy when replace the single MCT with the average operation.
  - Proving that the MCT can effectively fuse multimodal features by capturing the multimodal alignment.
- The influence of entity inconsistency is smaller.
  - Probably due to the sparsity of visual entities and the noises brought by entity detectors.

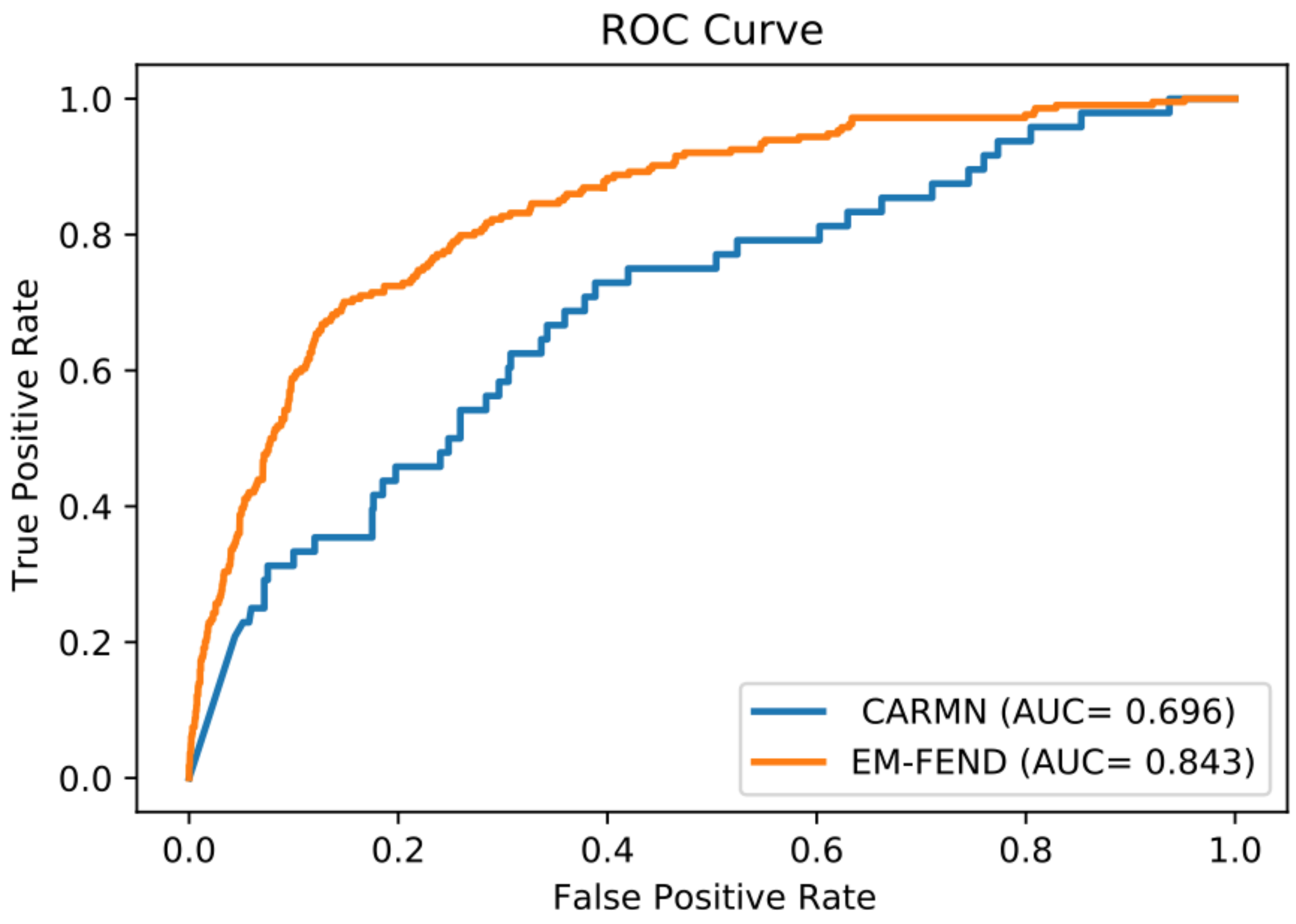
Experiments
Evaluation Questions

- EQ1: Can EM-FEND improve the classification performance of distinguishing multimodal fake and real news?
- EQ2: How effective are various visual features (especially visual entities) and cross-modal correlations in improving the performance of EM-FEND?
- EQ3: How does EM-FEND perform in online fake news detection?

Experiments

Robustness to Imbalance Data

- 217 fake : 3353 real = 1: 15
- Observe that EM-FEND outperform CARMN in online data.



Conclusion

- Find 3 valuable cross-modal correlations in multimodal FND on social media.
  - Entity inconsistency, mutual enhancement and text complementation.
- Reveal the importance of visual entities is in understanding news-related visual semantics and capturing these multimodal clues.
- Propose a novel entity-enhanced multimodal fusion framework named EM-FEND to simultaneously model 3 cross-modal correlations.

Comments
of EM-FEND

- Import concept of visual entities to align with textual feature.
- Visual entities API has limitation (?
  - People who didn't in the database?
- Didn't split two-type of visual feature (manipulated, non-manipulated).
  - If recognized the entities on manipulated picture?
  - May can fixed by attention mechanism?
- Recently approaches are used pre-trained LM (BERT) to encode textual feature.