Multimodal Fake News Detection via CLIP-Guided Learning

Yangming Zhou
School of Computer Science
Shanghai, China
ymzhou21@fudan.edu.cn

Qichao Ying
School of Computer Science
Shanghai, China
qcying20@fudan.edu.cn

Zhenxing Qian*
School of Computer Science
Shanghai, China
zxqian@fudan.edu.cn

Sheng Li
School of Computer Science
Shanghai, China
lisheng@fudan.edu.cn

Xinpeng Zhang
School of Computer Science
Shanghai, China
zhangxinpeng@fudan.edu.cn

Submitted to CIKM'22 (Under Preview)

220616 Chia-Chun Ho

Outline
of FND-CLIP

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction
Fake News Detection

- News forgery can take various forms
  - Replacing a critical object within a picture with another one
  - Making biased or even misleading comments on the picture
- Automatic FND using machine learning
  - An efficient way to combat the widespread dissemination of fake news
  - To help news readers identify bias and misinformation in news articles

Introduction
Unimodal FND

- Early works merely focused on text-only or image-only content analysis
- Unimodal FND schemes are effective
  - But modern news and posts are usually w/ rich information of several modalities
  - These method neglect their correlation
    - Multimodal feature analysis is required to offer complementary benefits to assist FND

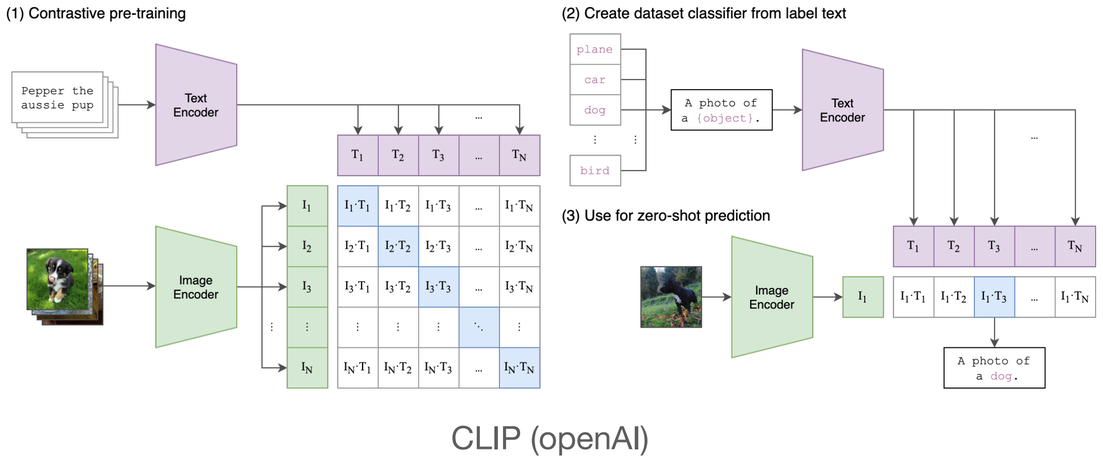
Introduction

Multimodal FND

- Fusing features from images & texts usually
- Comments, up-vote ratio & the spreading graph also be considered as reference
  - These additional modalities are interactive and change over time
  - Many previous works prefer using as much modalities as possible
  - However, interactive modalities are less dependable than images and texts
- Some works try to explicitly calculating correlation on generating fused features
  - CAFE, MVAE

Introduction

FND-CLIP



CLIP (openAI)

- Based on the pretrained Contrastive Language-Image Pretraining (CLIP) model
- Propose a multimodal fake news detector called FND-CLIP
  - To address the issue of cross-modal ambiguity
    - By explicitly measuring the correlation between texts and images of targeted posts
    - To guide the feature fusing and decision-making stages
  - Introduce an attention layer that outputs attention scores
    - Adaptively measure the significance of features in their contribution to fake news detection

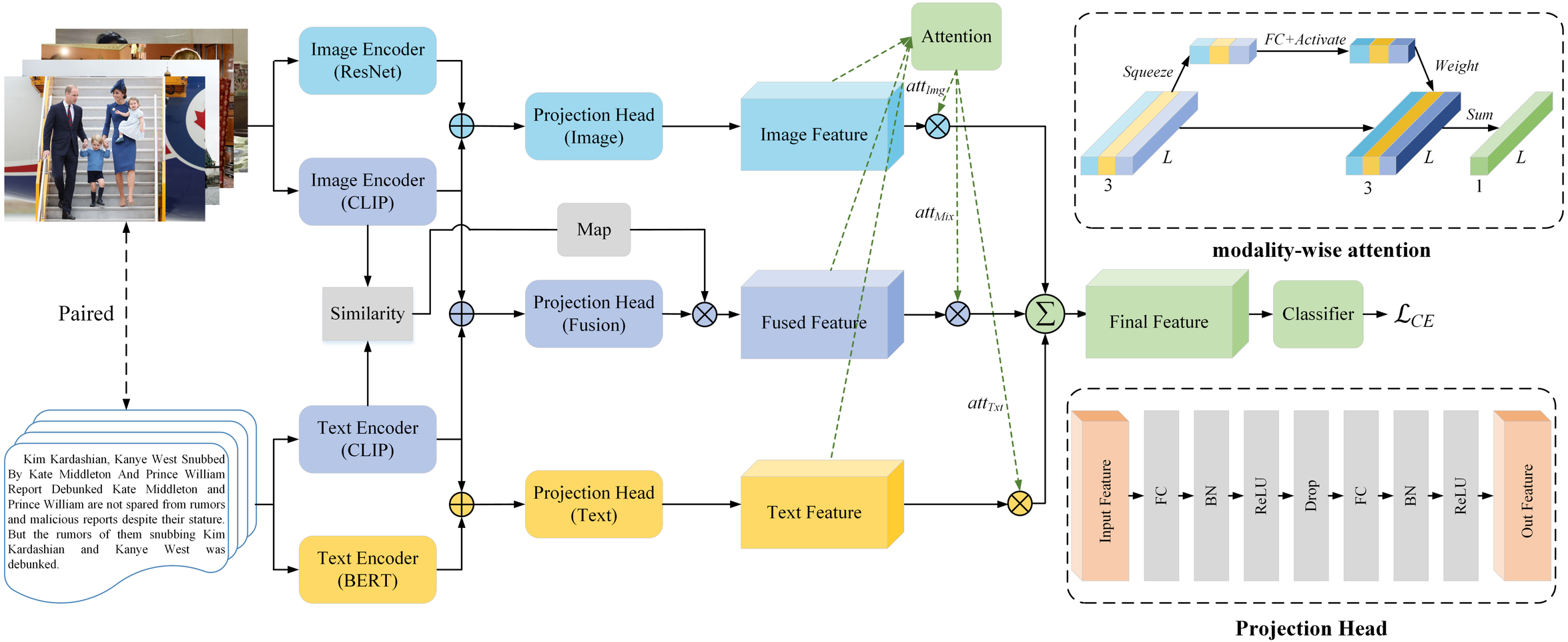
Introduction

Contributions

- Propose FND-CLIP, a multimodal FND method with CLIP-based learning
  - CLIP pretrained model is used to measure the cross-modal similarity and guide the mapping and fusion of features
- Propose a modality-wise attention mechanism
  - To adaptively weight the text, image, and fused features.
- Conducted comprehensive experiments on three datasets
  - Proving that CLIP-generated features can be important assists to the unimodal features

Methodology

FND-CLIP



Methodology

Unimodal Feature Generation

Paired

Kim Kardashian, Kanye West Snubbed By Kate Middleton And Prince William Report Debunked Kate Middleton and Prince William are not spared from rumors and malicious reports despite their stature. But the rumors of them snubbing Kim Kardashian and Kanye West was debunked.

Image Encoder (ResNet)

Image Encoder (CLIP)

Similarity

Text Encoder (CLIP)

Text Encoder (BERT)

f_{ResNet}

f_{CLIP-I}

f_{CLIP-T}

f_{BERT}

$f_{Img} = concat(f_{ResNet}, f_{CLIP-I})$

$f_{Txt} = concat(f_{BERT}, f_{CLIP-T})$

Methodology

CLIP-Guide Multimodal feature generation

Image Encoder (ResNet)

Image Encoder (CLIP)

Similarity

Paired

Kim Kardashian, Kanye West Snubbed By Kate Middleton And Prince William Report Debunked Kate Middleton and Prince William are not spared from rumors and malicious reports despite their stature. But the rumors of them snubbing Kim Kardashian and Kanye West was debunked.

Text Encoder (CLIP)

Text Encoder (BERT)

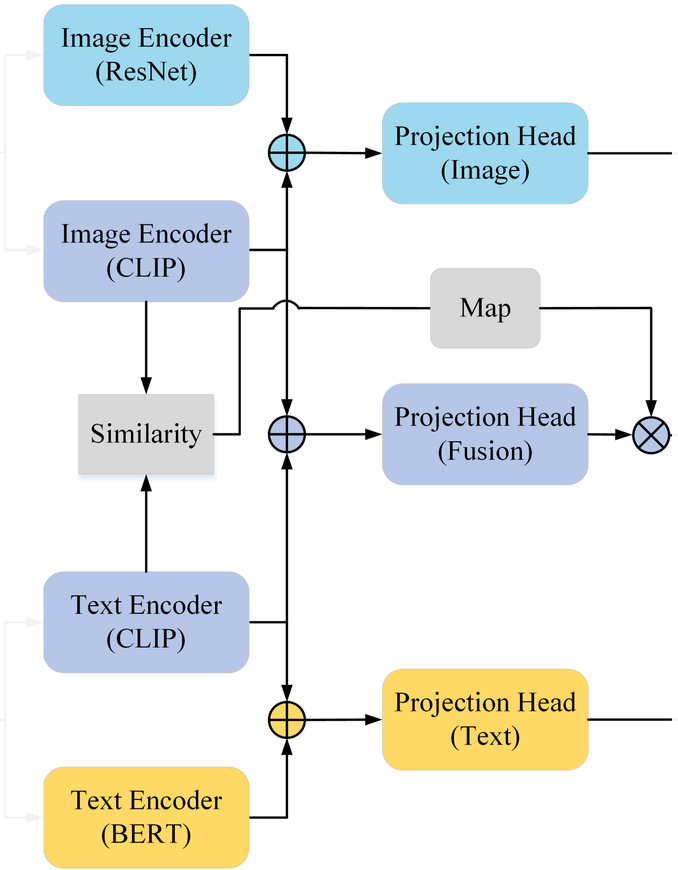
f_{CLIP-I}

f_{CLIP-T}

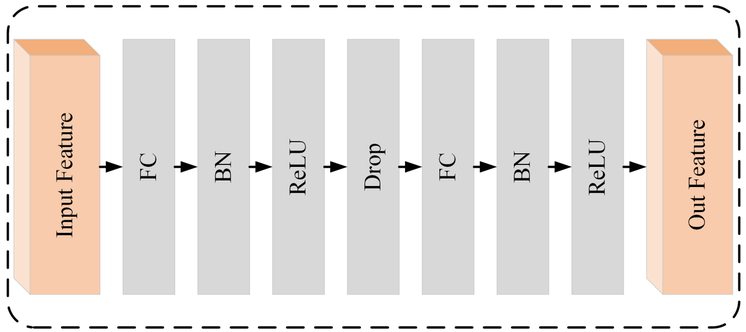
$$f_{Mix} = concat(f_{CLIP-T}, f_{CLIP-I})$$

Methodology

Projection Heads



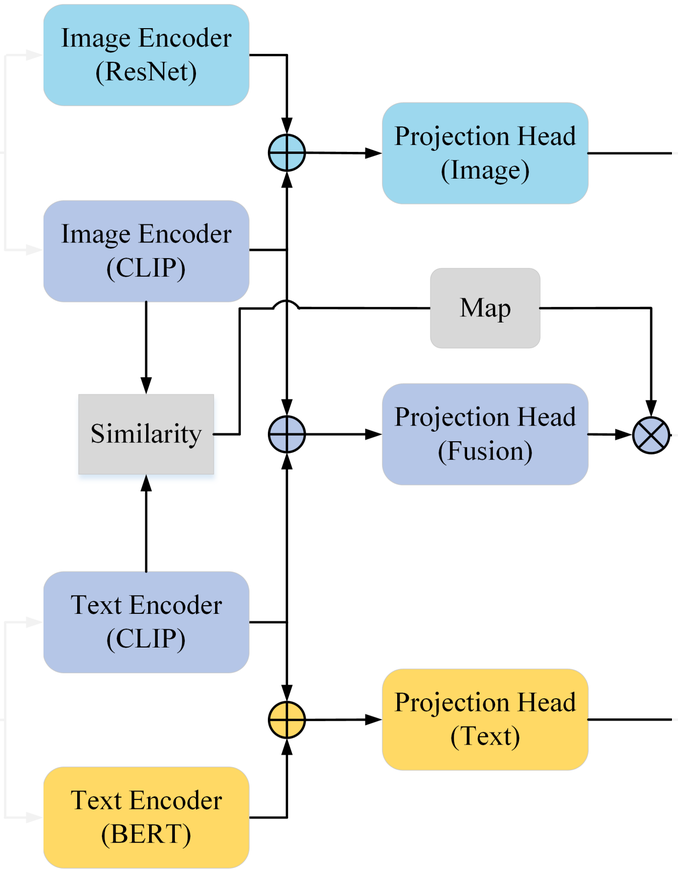
- To reduce the dimension of the coarse features provided by the encoders
  - Help filtering out redundant information
- Share same arch but do not share weight



Projection Head

Methodology

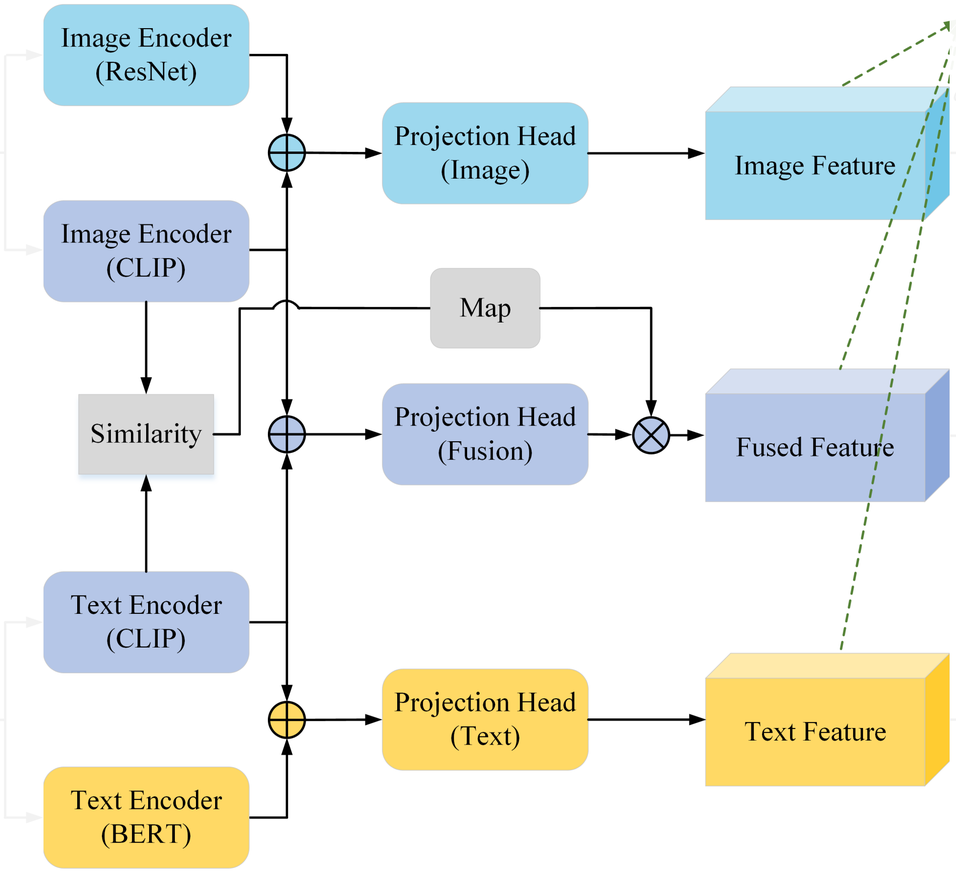
Ambiguity Between Modalities



- Merely combining the CLIP-based features cannot necessarily provide enough reliable information
- To address the ambiguity issue between multimodal features, measure the cosine similarity
  - $\text{sim} = \dfrac{f_{Txt} \cdot (f_{Img})^T}{\|f_{Txt}\| \, \|f_{Img}\|}$ (formula in paper)
  - $\text{sim} = \dfrac{f_{CLIP-T} \cdot (f_{CLIP-I})^T}{\|f_{CLIP-T}\| \, \|f_{CLIP-I}\|}$ (actual)

Methodology

Projected unimodal and multimodal features



$$m_{Img} = P_{Img}(f_{Img})$$

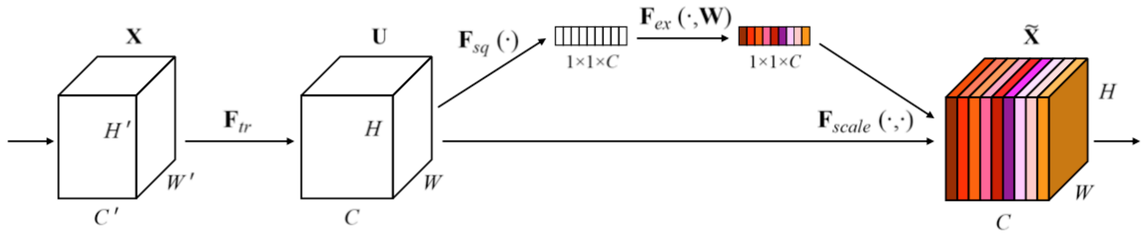
$$m_{Mix} = Sigmoid(Std(sim)) \cdot P_{Mix}(f_{Mix})$$

$$m_{Txt} = P_{Txt}(f_{Txt})$$

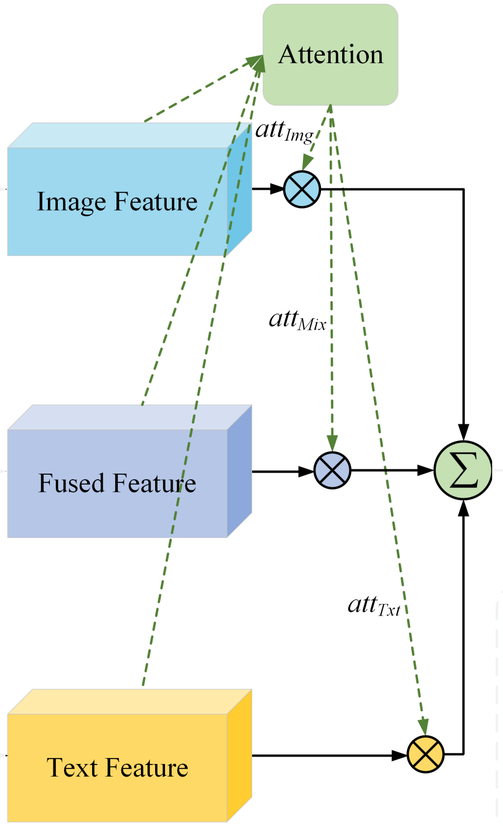
Methodology

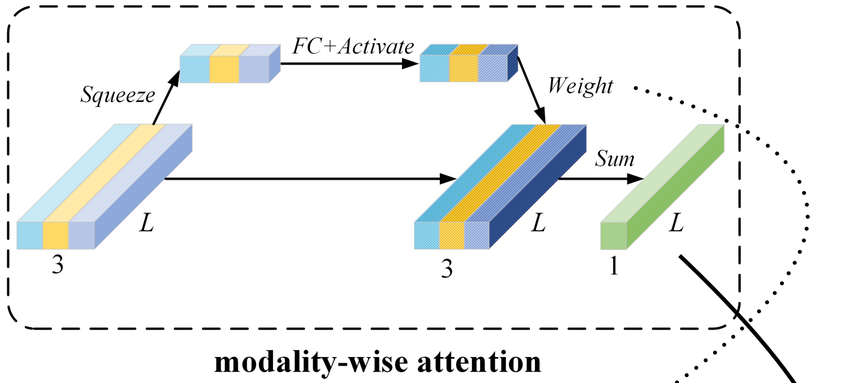
Projected unimodal and multimodal features

- Apply an attention mechanism to reweigh the projected features
- Inspired by the Squeeze-and-Excitation Networks (SE-Net)
  - Designed a modality-wise attention module to weight each feature adaptively



SE-Net (CVPR'18)





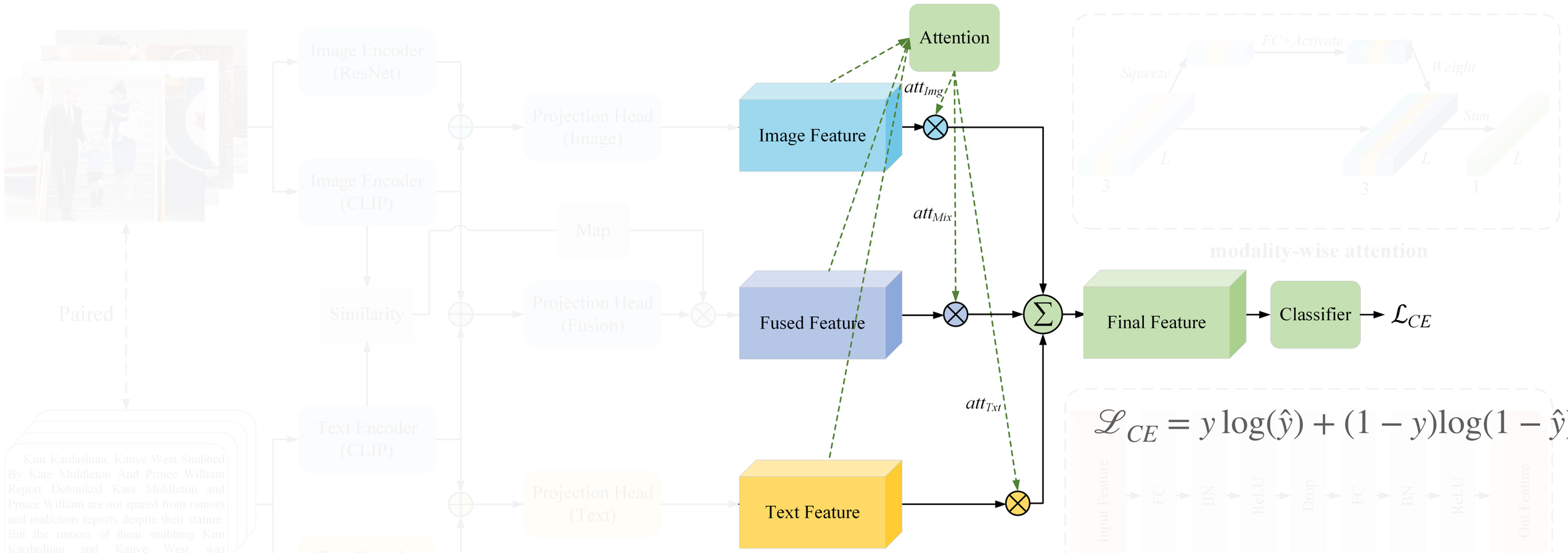
modality-wise attention

$$att = \{att_{Txt}, att_{Img}, att_{Mix}\}$$

$$m_{Agg} = att_{Txt} \cdot m_{Txt} + att_{Img} \cdot m_{Img} + att_{Mix} \cdot m_{Mix}$$

Methodology

Classification & Objective Function

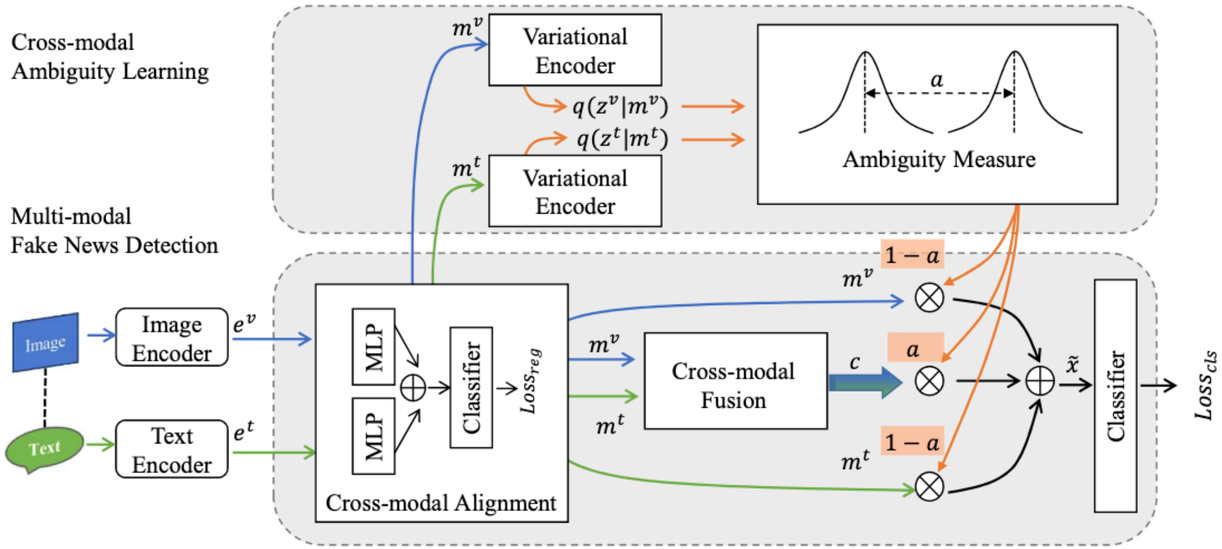


Experiments

Datasets & Baselines

	Weibo	PolitiFact	GossipCop
Train (Real, Fake)	3749, 3783	244, 135	7974, 2036
Test (Real, Fake)	1996*	75, 29	2285, 545

- EANN
- MVAE
- SpotFake
- MVNN
- SAFE
- LIIMR
- MCAN
- RoBERTa-MWSS
- SpotFake+
- TM
- LSTM-ATT
- DistilBERT
- CAFE



Experiments

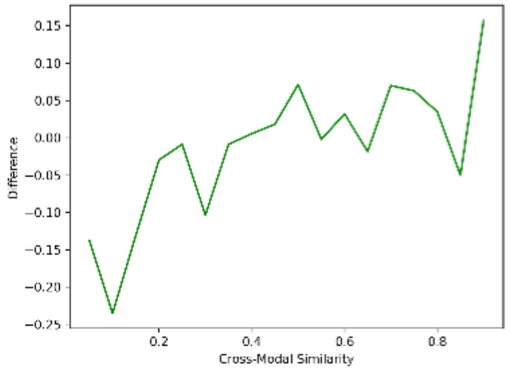
Performance Analysis

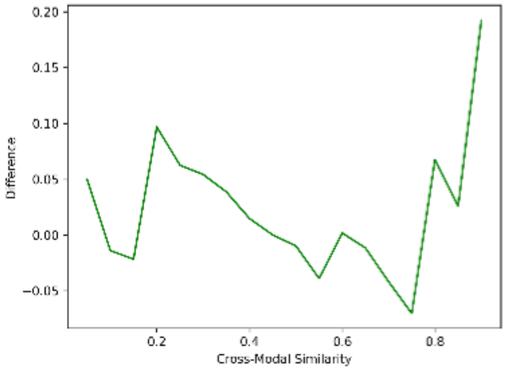
	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	EANN [42]	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	MVAE [24]	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	Spotfake [40]	0.892	0.902	**0.964**	**0.932**	0.847	0.656	0.739
	MVNN [45]	0.846	0.809	0.857	0.832	0.879	0.837	0.858
	SAFE [48]	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	LIIMR [39]	0.900	0.882	0.823	0.847	0.908	**0.941**	**0.925**
	MCAN [44]	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	CAFE [9]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP	**0.907**	**0.914**	0.901	0.908	**0.914**	0.901	0.907
Politifact	RoBERTa-MWSS [37]	0.820		-	-	0.820	-	-
	SAFE [48]	0.874	0.851	0.830	0.840	0.889	0.903	0.896
	Spotfake+ [38]	0.846	-	-	-	-	-	-
	TM [4]	0.871	-	-	-	0.901	-	-
	LSTM-ATT [27]	0.832	0.828	0.832	0.830	0.836	0.832	0.829
	DistilBert [1]	0.741	0.875	0.636	0.737	0.647	0.880	0.746
	CAFE [9]	0.864	0.724	0.778	0.750	0.895	0.919	0.907
	FND-CLIP	**0.942**	**0.897**	**0.897**	**0.897**	**0.960**	**0.960**	**0.960**
Gossipcop	RoBERTa-MWSS [37]	0.800	-	-	0.800	-	-	-
	SAFE [48]	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	Spotfake+ [38]	0.856	-	-	-	-	-	-
	TM [4]	0.842	-	-	-	0.896	-	-
	LSTM-ATT [27]	0.842	**0.845**	**0.842**	**0.844**	0.839	0.842	0.821
	DistilBert [1]	0.857	0.805	0.527	0.637	0.866	**0.960**	0.911
	CAFE [9]	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	FND-CLIP	**0.880**	0.761	0.549	0.638	**0.899**	0.959	**0.928**

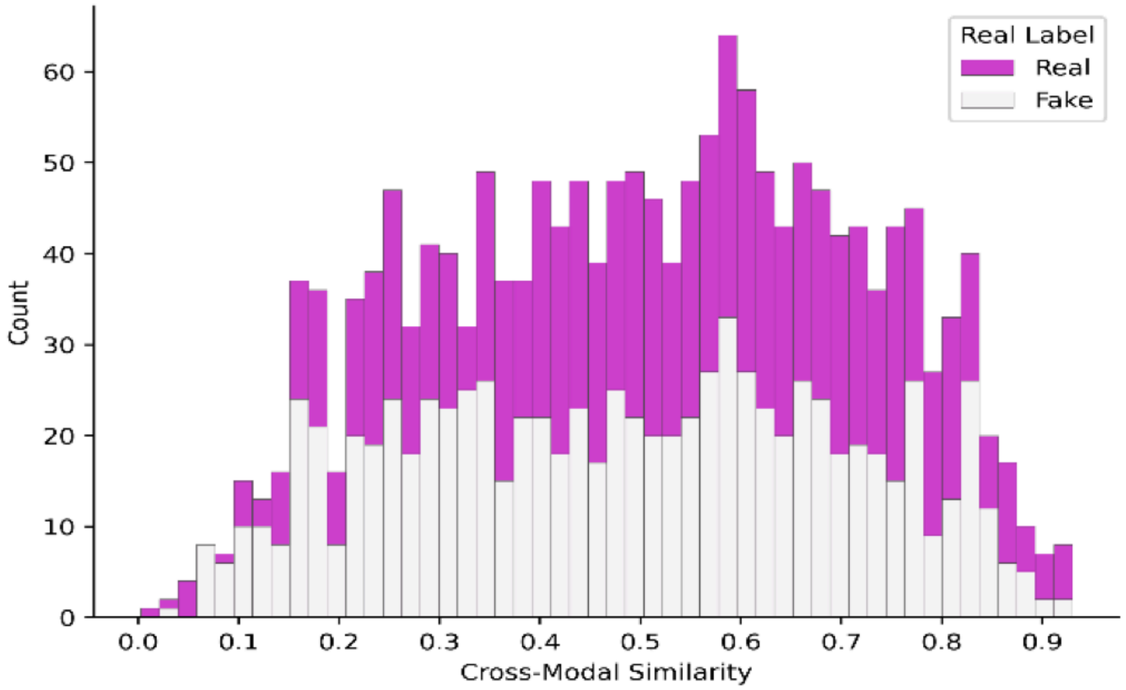
	Weibo	PolitiFact	GossipCop
Train (Real, Fake)	3749, 3783	244, 135	7974, 2036
Test (Real, Fake)	1996*	75, 29	2285, 545

Experiments

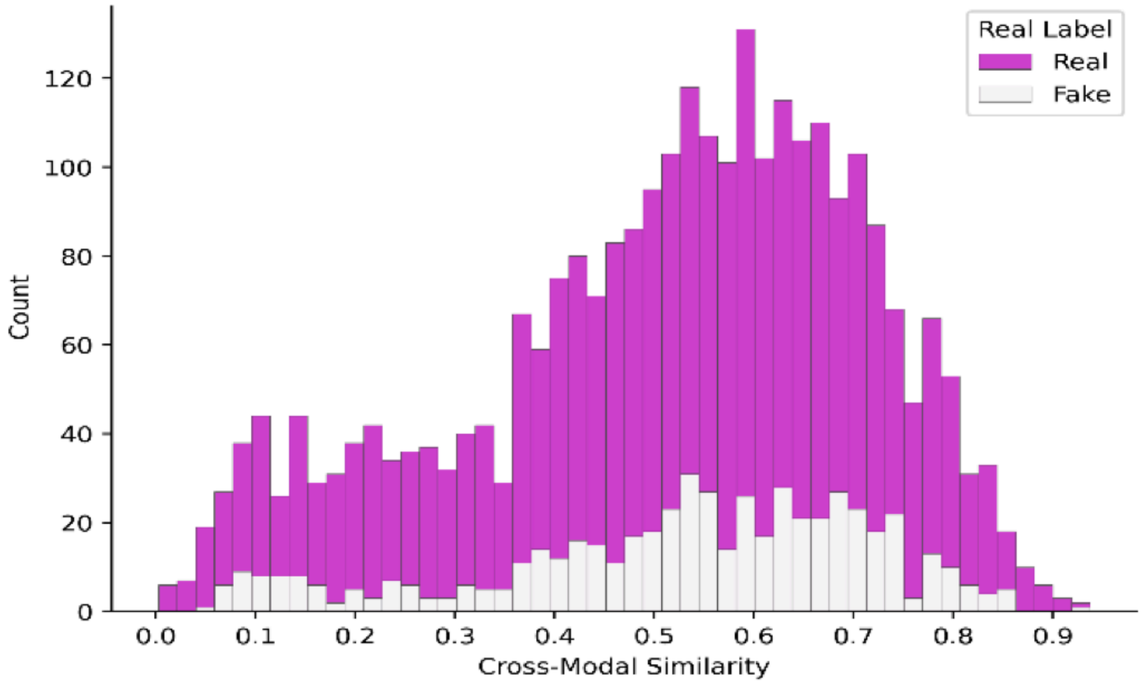
Statistical Analysis







Weibo

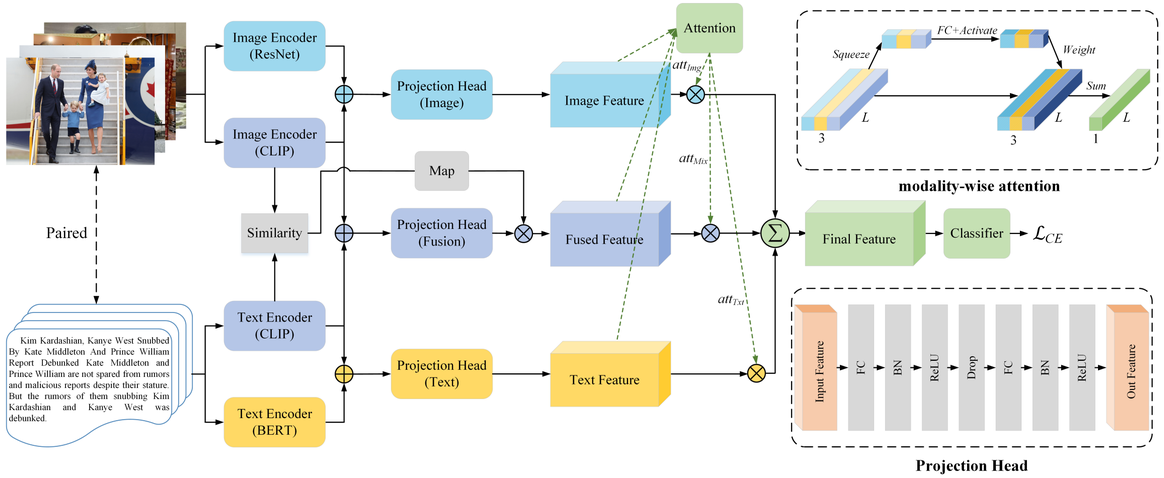


GossipCop

Experiments

Ablation Study

- w/o C (CLIP-related modules)
  - Only use BERT & ResNet to extract features
- w/o F (Fusion module)
  - Use two unimodal features to classify
- w/o A (Attention module)
  - Direct aggregate the three features to obtain final feature

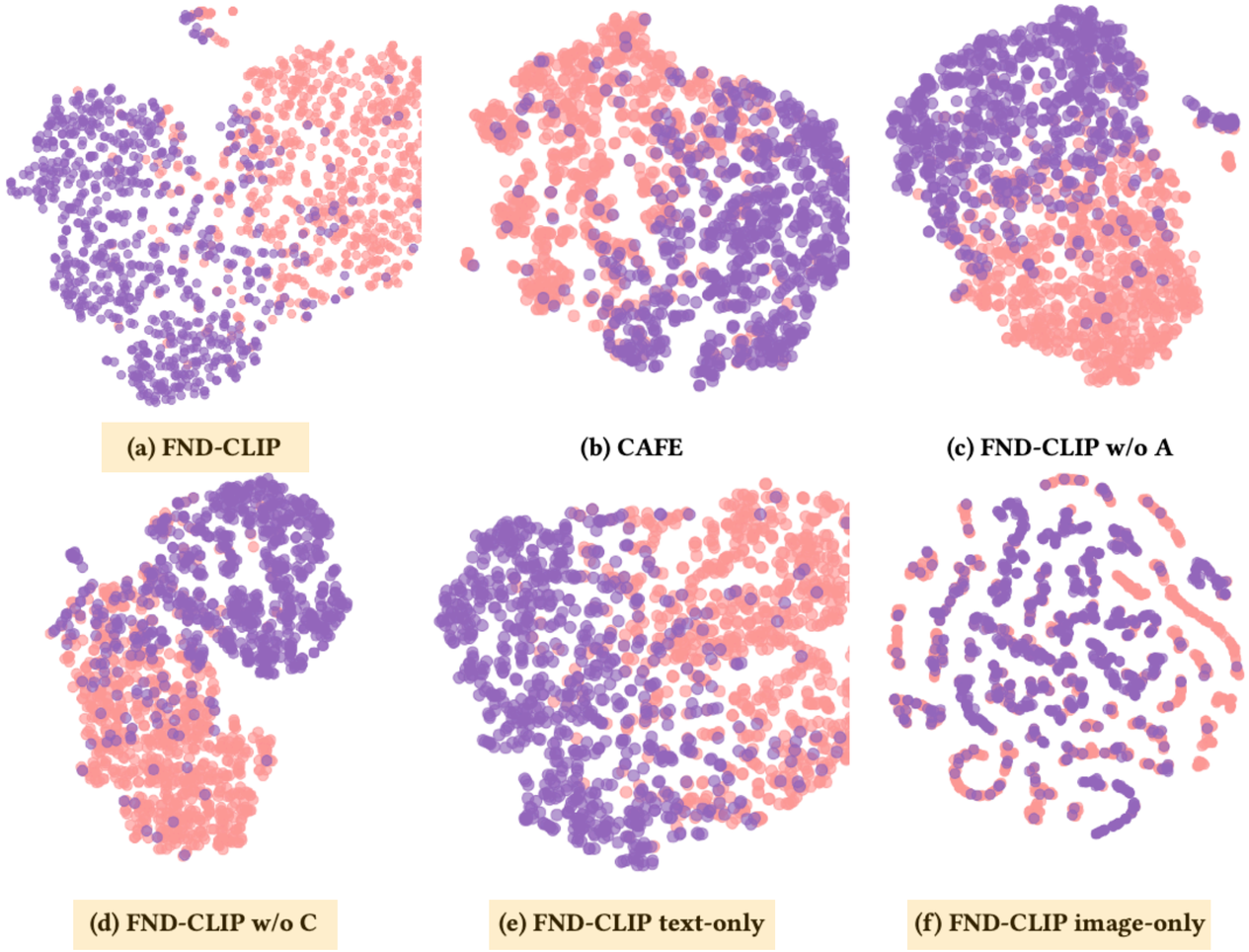


	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	FND-CLIP multimodal-only	0.817	0.899	0.718	0.798	0.761	0.917	0.832
	FND-CLIP image-only	0.796	0.862	0.711	0.779	0.750	0.884	0.811
	FND-CLIP text-only	0.872	0.906	0.833	0.868	0.842	0.911	0.875
	FND-CLIP w/o C	0.874	0.895	0.851	0.872	0.855	0.898	0.876
	FND-CLIP w/o F	0.893	0.925	0.857	0.890	0.864	0.929	0.895
	FND-CLIP w/o A	0.897	**0.936**	0.855	0.893	0.863	**0.940**	0.900
	FND-CLIP	**0.907**	0.914	**0.901**	**0.908**	**0.901**	0.914	**0.907**
Politifact	FND-CLIP multimodal-only	0.903	0.807	0.862	0.833	0.944	0.919	0.932
	FND-CLIP image-only	0.748	0.600	0.310	0.409	0.773	0.919	0.840
	FND-CLIP text-only	0.903	0.913	0.724	0.808	0.900	**0.973**	0.935
	FND-CLIP w/o C	0.893	0.875	0.724	0.793	0.899	0.960	0.928
	FND-CLIP w/o F	0.903	0.880	0.759	0.815	0.910	0.960	0.934
	FND-CLIP w/o A	**0.942**	**0.926**	0.862	0.893	0.947	**0.973**	**0.960**
	FND-CLIP	**0.942**	0.897	**0.897**	**0.897**	**0.960**	0.960	**0.960**
Gossipcop	FND-CLIP multimodal-only	0.862	0.708	0.484	0.575	0.886	0.952	0.918
	FND-CLIP image-only	0.814	**1.000**	0.033	0.064	0.813	**1.000**	0.897
	FND-CLIP text-only	0.871	0.741	0.508	0.603	0.891	0.958	0.923
	FND-CLIP w/o C	0.870	0.745	0.494	0.594	0.888	0.960	0.923
	FND-CLIP w/o F	0.874	0.723	0.562	0.632	0.901	0.949	0.924
	FND-CLIP w/o A	0.873	0.715	**0.567**	0.633	**0.902**	0.946	0.923
	FND-CLIP	**0.880**	0.761	0.549	**0.638**	0.899	0.959	**0.928**

Experiments

t-SNE Visualizations

- w/o C (CLIP-related modules)
  - Only use BERT & ResNet to extract features
- w/o F (Fusion module)
  - Use two unimodal features to classify
- w/o A (Attention module)
  - Direct aggregate the three features to obtain final feature



t-SNE on Weibo

Conclusion
of FND-CLIP

- Proposed novel FND method called FND-CLIP
  - Use CLIP to extract aligned multimodal features & guide the learning of network for different modalities
  - Introduce modality-wise attention to adaptively determine the weights of text, image and fused features
    - Avoid introducing noisy and redundant features during features
    - Further improve the classification accuracy

Comments
of FND-CLIP

- Focus on cross-modal ambiguity
  - Calculate the modal similarity
- Usage of the similarity is different w/ CAFE
  - Similarity high use more fused feature
- GossipCop can consider as improve goal
  - F1, Precision, Recall