Cross-modal Ambiguity Learning for Multimodal Fake News Detection

Yixuan Chen
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
yixuanchen20@fudan.edu.cn

Dongsheng Li
Microsoft Research Asia
Shanghai, China
dongsli@microsoft.com

Peng Zhang
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
zhangpeng_@fudan.edu.cn

Jie Sui
University of Chinese Academy of Sciences
Beijing, China
suijie@ucas.ac.cn

Qin Lv
University of Colorado Boulder
Boulder, United States
qin.lv@colorado.edu

Tun Lu, Li Shang*
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
{lutun,lishang}@fudan.edu.cn

WWW'22

220519 Chia-Chun Ho

Outline
of CAFE

Introduction

Methodology

Experiments

Conclusion

Comments

Introduction

Fake News on Social Media

- Online social media has become the primary information platform for people.
- Over three billion people consider Facebook & Twitter as their primary daily information sources.
- Lack of systematic efforts to verify the credibility of online posts.
  - Led to the wide and fast spread of fake news across social platforms.
- Fake news detection has received increasing research attention in recent years.

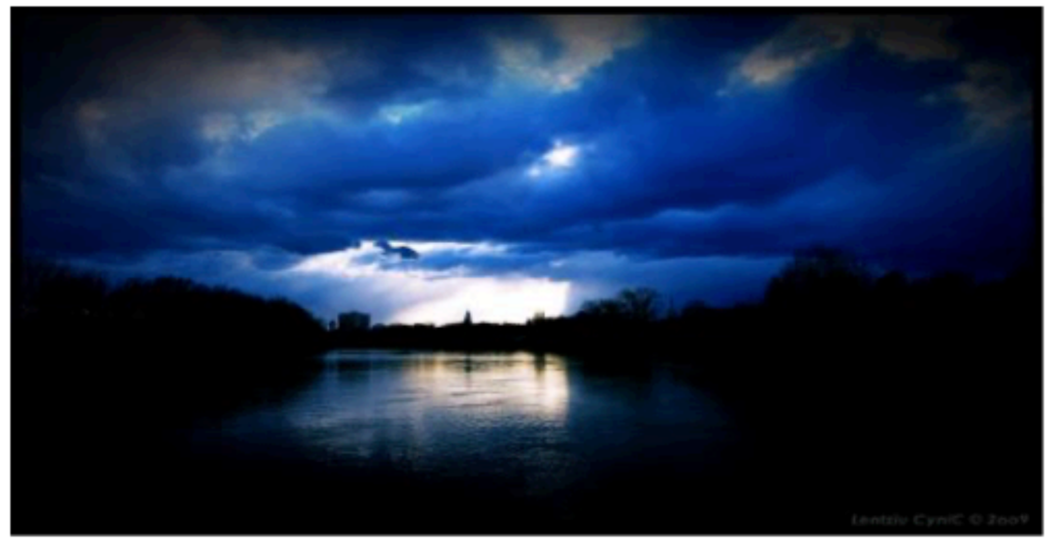
Introduction
Multi-modal content

- Online social content has quickly evolved from text-only to multimodal (text, image).
- Early works on fake news detection focus on text-only analysis.
  - Cross-modal analysis can offer complementary benefits to assist the FND task.
- Recent works aim to fuse multimodal information to boost performance.
- However, prior works have not explicitly considered the inherent ambiguity across different modalities.
  - Thus lead to inferior performance.

Introduction

Multimodal Example

Fake news: "An employee of the Jefferson County morgue died this morning after being accidentally cremated by one of his coworkers."



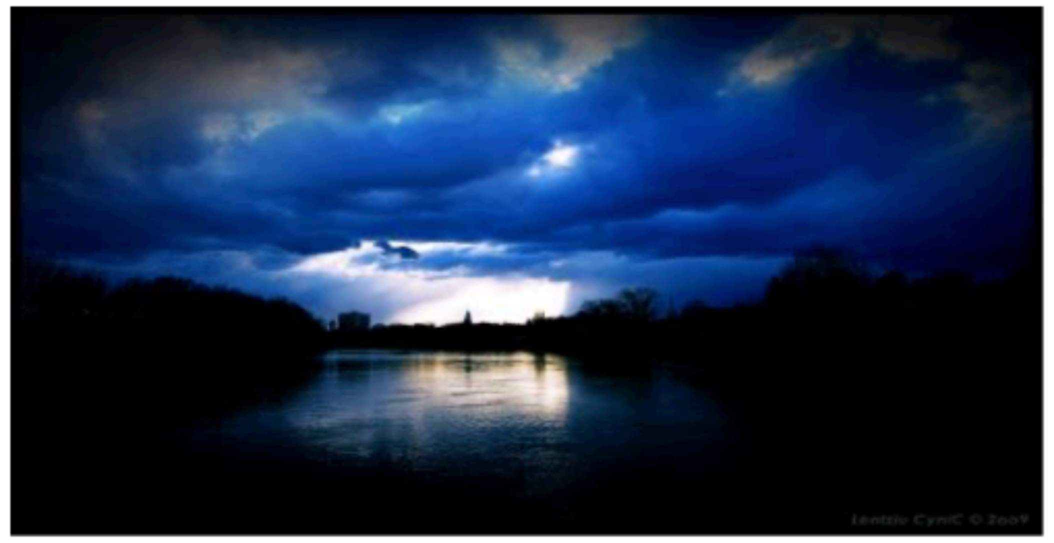
Real news: "You left in peace, left me in pieces."

- Fake news example (left figure) tells a fictional death story but includes a smiling person's image.
  - Text & image present strong cross-modal ambiguity.
  - Multimodal feature fusion captures such cross-modal information gap.
    - Help improve classification accuracy.

Introduction

Multimodal Example

Fake news: "An employee of the Jefferson County morgue died this morning after being accidentally cremated by one of his coworkers."



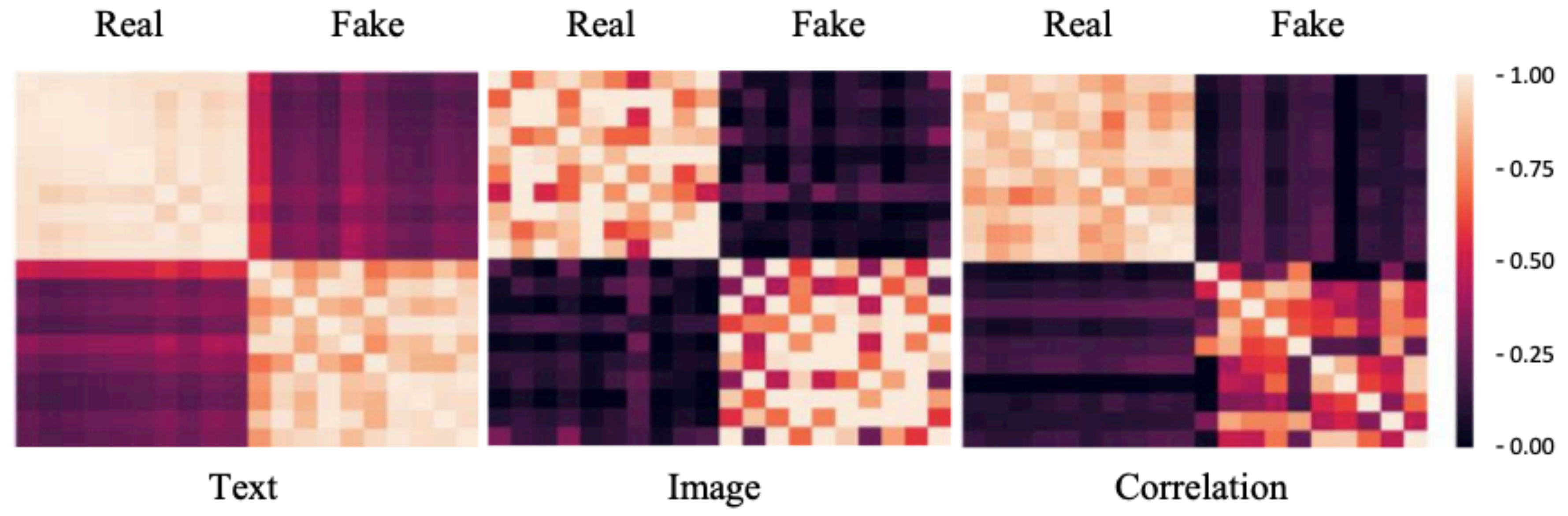
Real news: "You left in peace, left me in pieces."

- In contrast, real news example (right figure) expresses sad emotion with a blue image included.
  - Uni-modalities are emotionally consistent and are sufficient to determine news credibility.
  - Cross-modal fusion features are unnecessary or even introduce noise to the classification task.

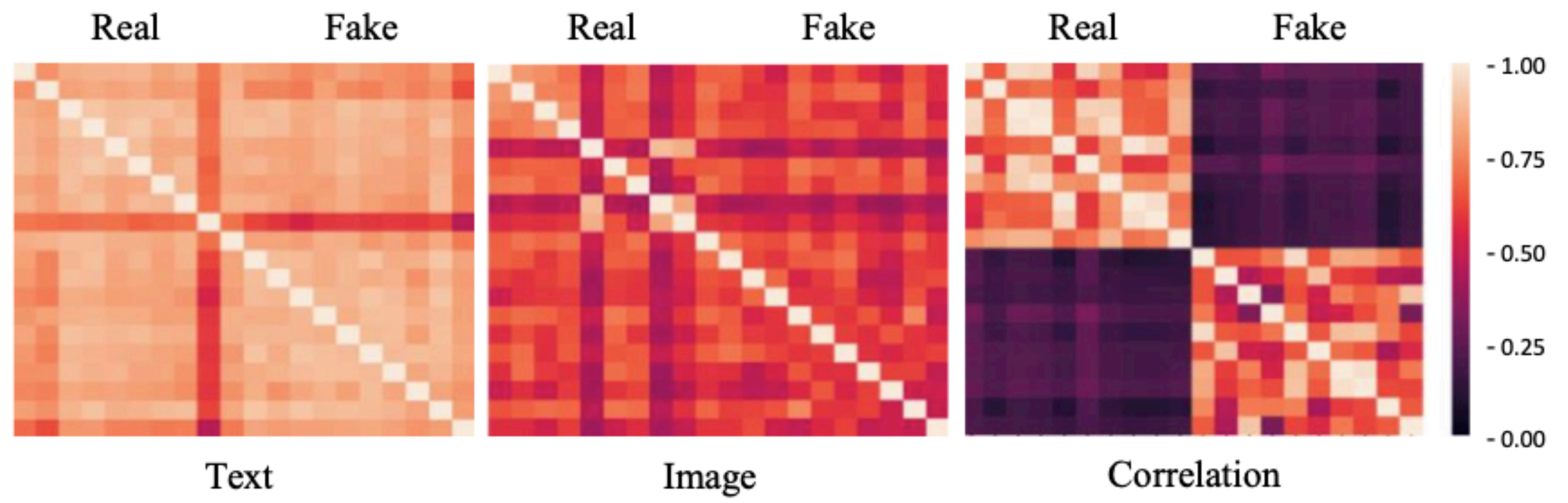
Introduction

Statistical Analysis

- Cross-modal information may be unhelpful or even harmful when unimodal are sufficient and agree with each other.
- Cross-modal information is crucial when unimodal are insufficient.
- Methods should be aware of the ambiguity between different modalities and adaptively aggregate discriminative cross-modal features with unimodal features.



(a) Cross-modal correlation may be unhelpful or even harmful when text and image alone are sufficient.



(b) Cross-modal correlation can present extra insights when text and image alone are insufficient.

(a) 42.9% posts (b) 11.9% of Weibo dataset

Introduction

CAFE (Cross-modal ambiguity-aware multimodal fake news detection)

- Formulate the cross-modal ambiguity learning problem in this paper.
  - By using the distributional divergence between different unimodal features to quantify their ambiguity.
- Propose CAFE – an ambiguity-aware multimodal fake news detection method.
  - Transform the heterogeneous unimodal features into a shared semantic space.
  - Estimate the ambiguity between different modalities.
  - Capture the cross-modal correlations by learning the semantic interactions between different modalities.

Introduction

CAFE (Cross-modal ambiguity-aware multimodal fake news detection)

- CAFE can improve fake news detection accuracy by adaptively aggregating unimodal features and cross-modal correlations.
    - Relying on unimodal features when cross-modal ambiguity is weak.
    - Referring to cross-modal correlations when cross-modal ambiguity is strong.

Introduction

Contributions

- Formulate the cross-modal ambiguity learning problem.
  - Key challenge to multimodal fake news detection.
  - Present a KL divergence based method to quantify the ambiguity between text and image by estimating the divergence of their feature distributions.
- Propose CAFE - an ambiguity-aware multimodal fake news detection method.
  - Adaptively aggregate unimodal features and cross-modal correlations, governed by the learnt ambiguity score.

Methodology

Problem Definition

- Formulate the key problem to multimodal fake news detection.
  - Cross-modal ambiguity learning

cross-modal ambiguity
between i-th & j-th modality

collection of information from all n modalities

$$a_{\mathbf{x}}^{i,j} = D(x^i, x^j), (\mathbf{x}, y) = \{\{x^u\}_n, y\}$$

distance measure function

label of $\mathbf{x}$

denotes information
from u-th modality
could be text, image, video, etc.

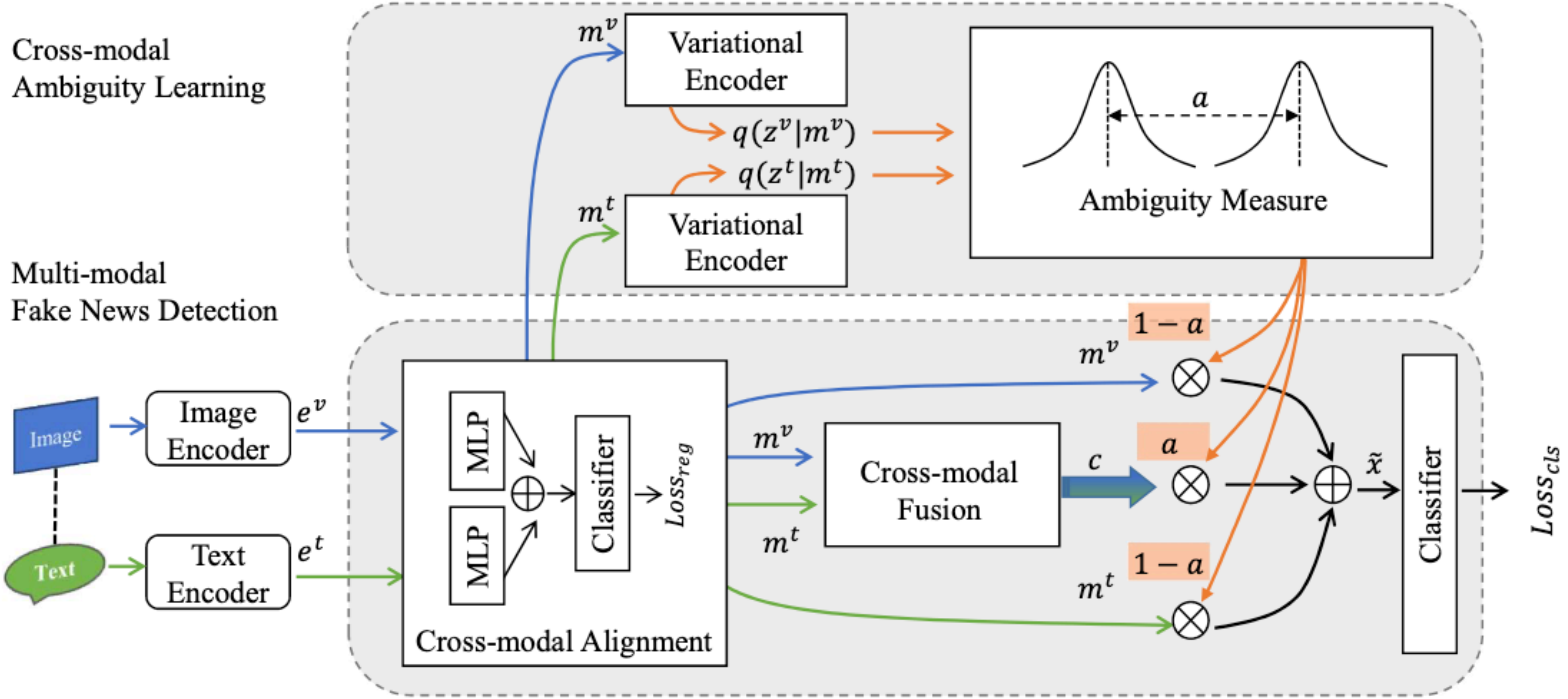
Methodology

Cross-modal ambiguity

- Measures the information gap between unimodal information.
- Cross-modal ambiguity learning is an important measure to decide when unimodal information is sufficient and when cross-modal information is essential.
  - It can be measured by the similarity, or the distance, between unimodal distributions using methods such as KL divergence and Wasserstein distance.
  - In this work, propose a KL divergence based method for cross-modal ambiguity learning.

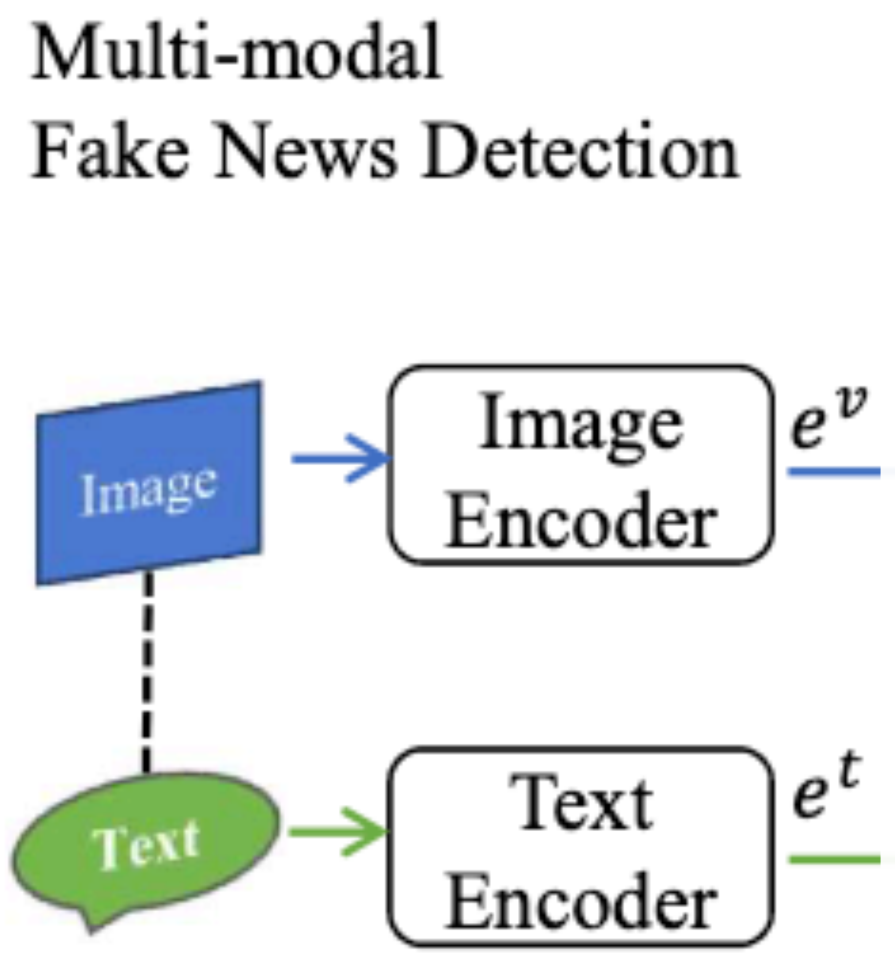
Methodology

CAFE (Cross-modal ambiguity-aware multimodal fake news detection)



Methodology

Modal-specific Encoder

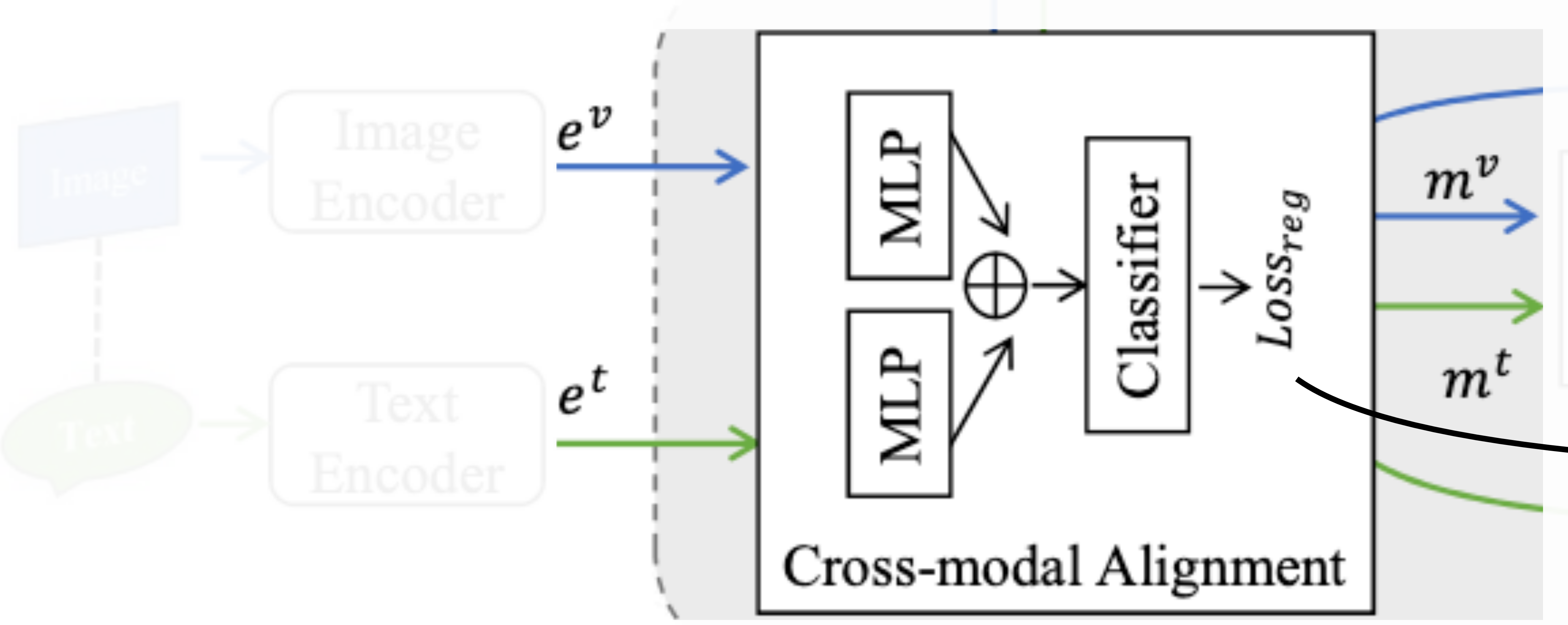


- Since the modal-specific encoders are not the focus of this work, adopt the off-the-shelf techniques.
  - Text Encoder
    - Adopt pretrained BERT model to obtain its embedding e^t.
  - Image Encoder
    - Adopt popular pretrained method – ResNet-34 to learn meaningful representations e^v from images.

Methodology

Cross-modal Alignment

- Features may have huge semantic gaps.
  - Need to align the features from different modalities by transforming the unimodal embeddings into a shared space.



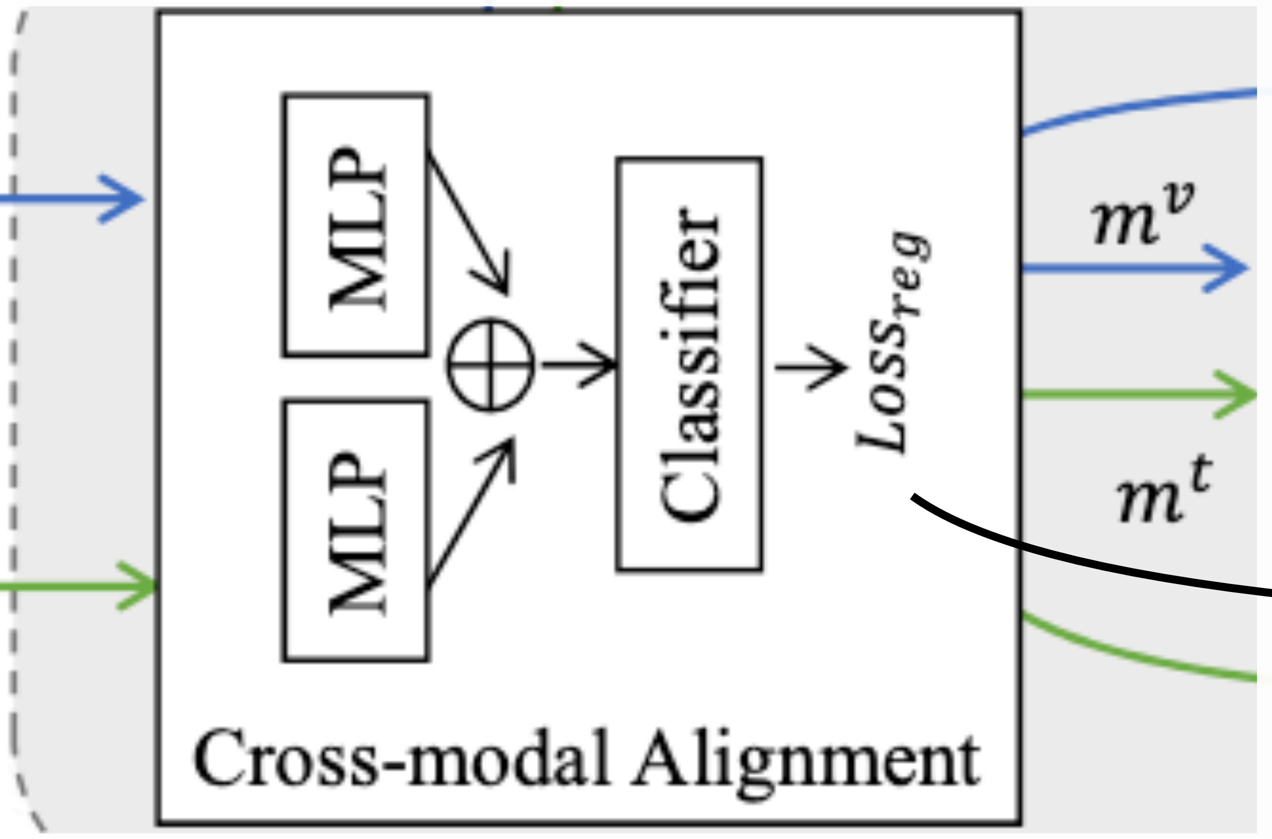
$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } y_2 = 0 \end{cases}$$

Methodology

Cross-modal Alignment

- Propose to solve an auxiliary correlation learning task to help achieve cross-modal feature alignment.

- Design a binary classification task to identify whether a pair of textual and visual embeddings shares a common semantics or not. (Semantic Regularization)



$$\mathscr{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } y_2 = 0 \end{cases}$$

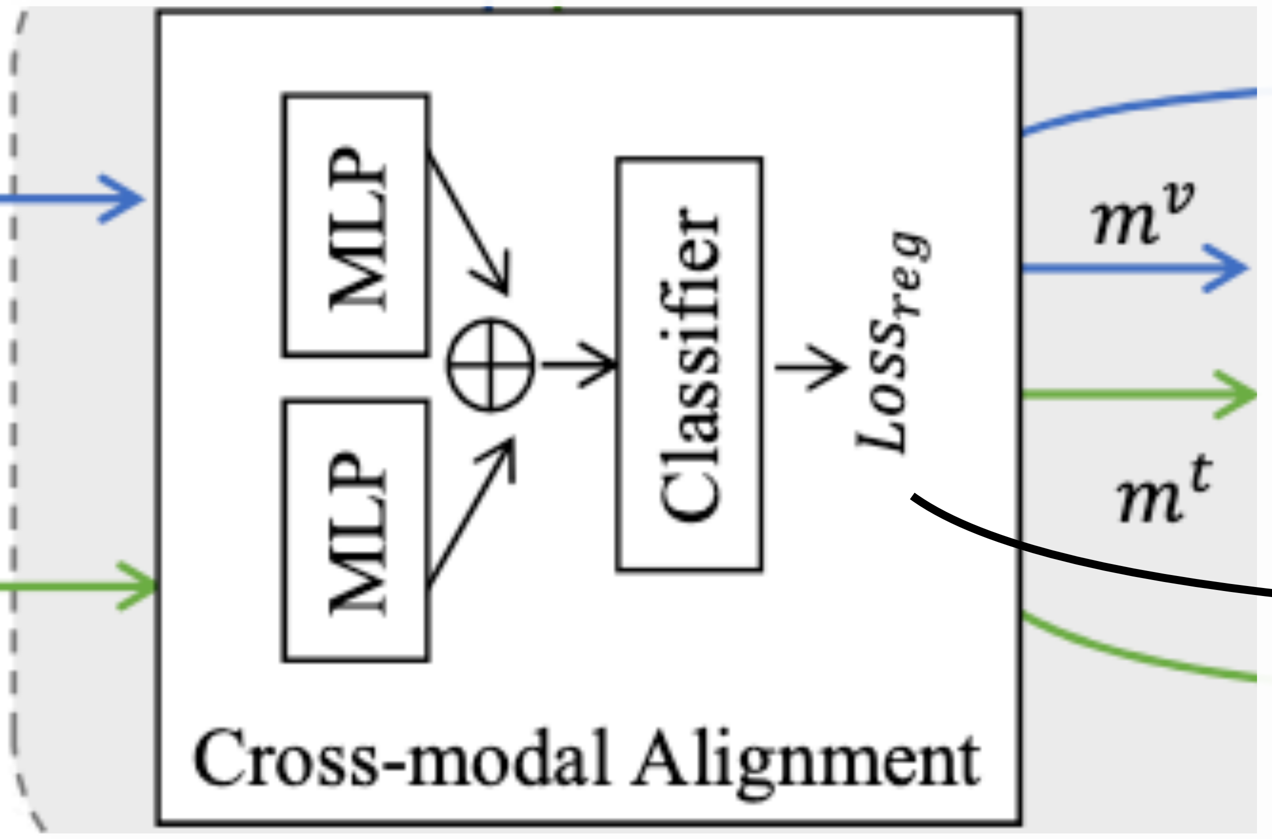
Methodology

Cross-modal Alignment

- Given each text-image pair, first define the semantic correlation is positive or negative, i.e., labeled by 1 or 0.
- Semantic correlation of a text-image pair is defined as
  - Positive if the textual and visual embeddings are from the same piece of real news.
  - Negative if the textual and visual embeddings are from different pieces of real news.
- Randomly sample positive text-image pairs and negative text-image pairs to generate a synthetic dataset for the auxiliary correlation learning task.

Methodology
Cross-modal Alignment

- Proposed cross-modality alignment module consists of a modality-specific multilayer perceptron (MLP) and a modality-shared layer to jointly learn the shared semantics.
- Then, the joint embeddings are fed to an average pooling layer, which is followed by a fully connected layer as a binary classifier.



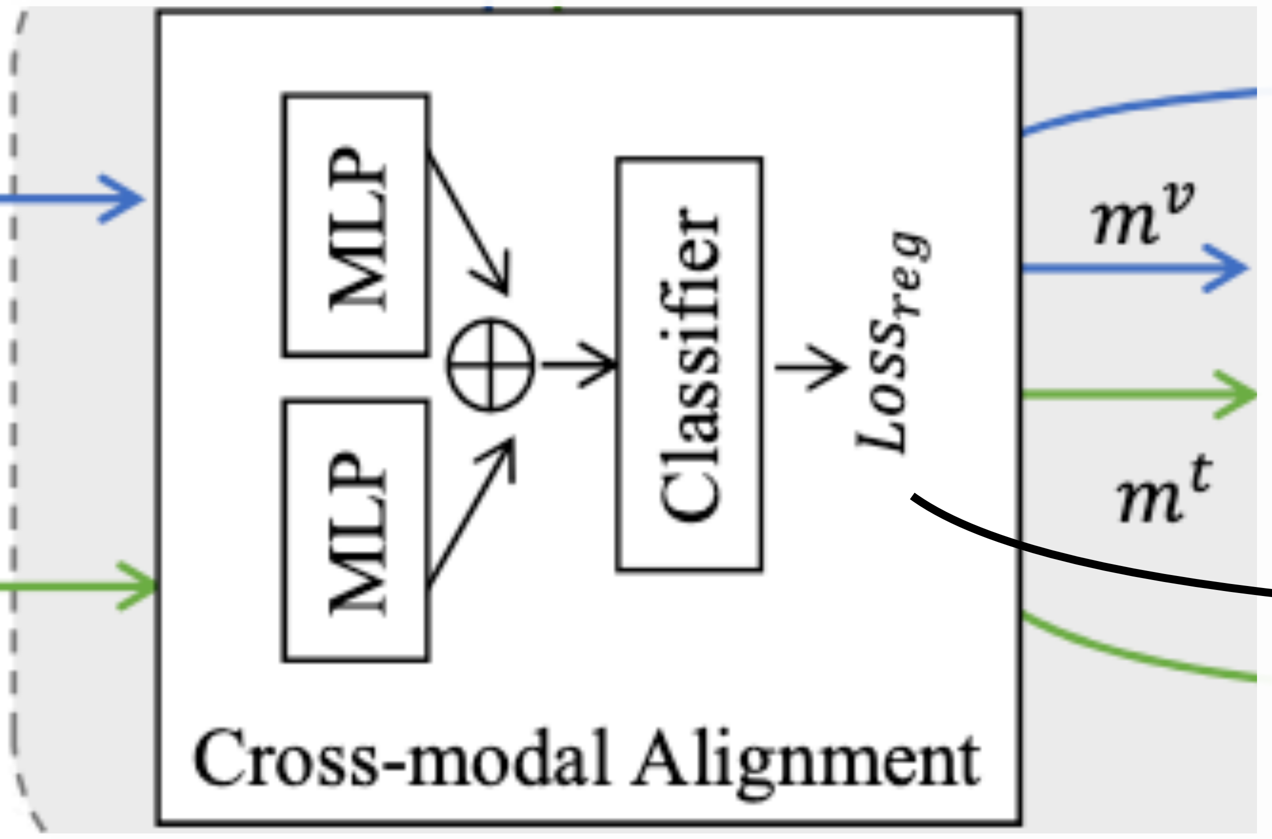
- The entire module is trained with positive & negative pairs using cosine embedding loss with margin d as follows:

$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } y_2 = 0 \end{cases}$$

Methodology

Cross-modal Alignment

- Above objective is to maximize the cosine similarity of embeddings between positive text-image pairs, and minimize it between negative pairs, up to a specified margin.
- With gradients from back-propagation, semantic regularization can automatically force heterogeneous multimodal embeddings into a shared semantic space.

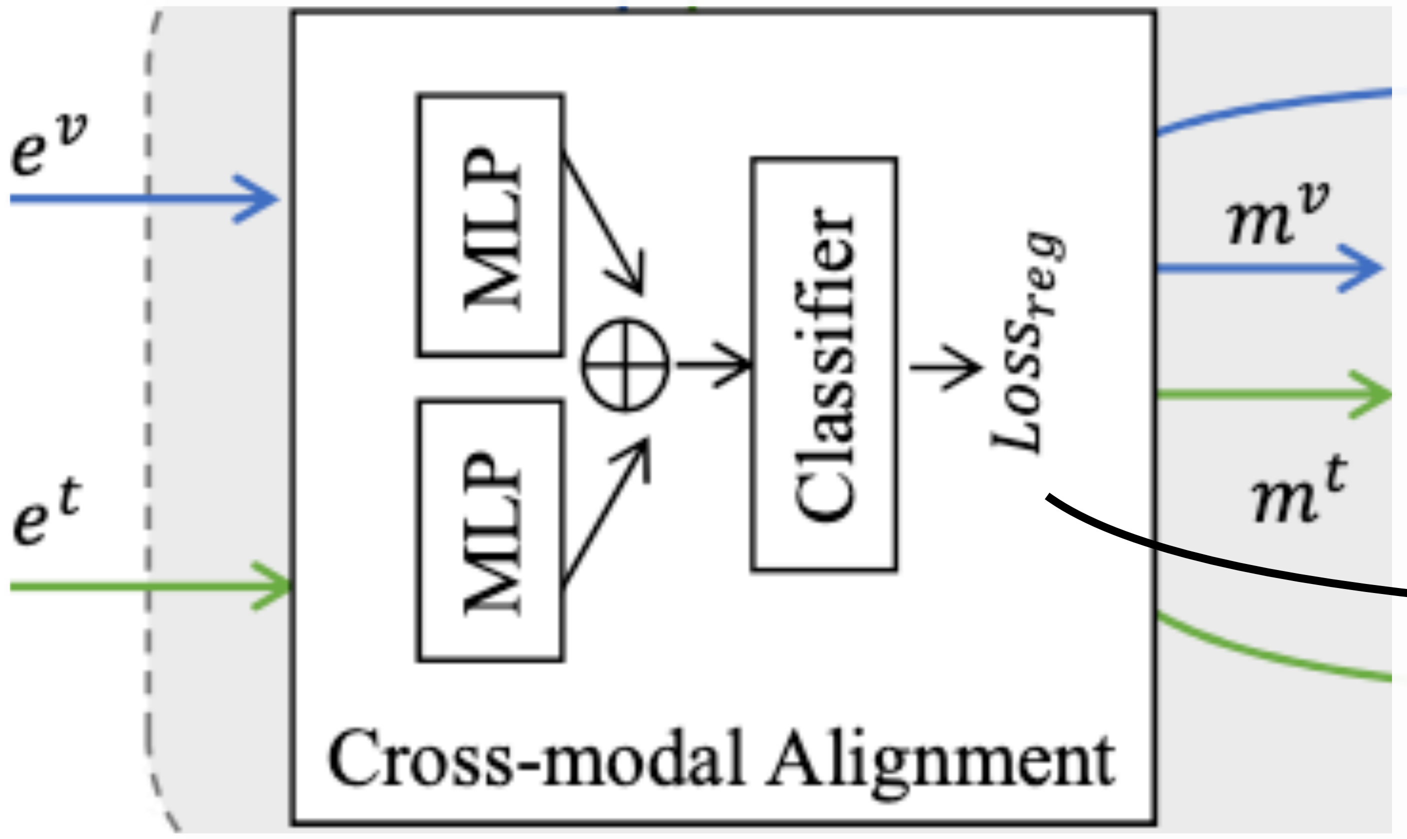


$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } y_2 = 0 \end{cases}$$

Methodology

Cross-modal Alignment

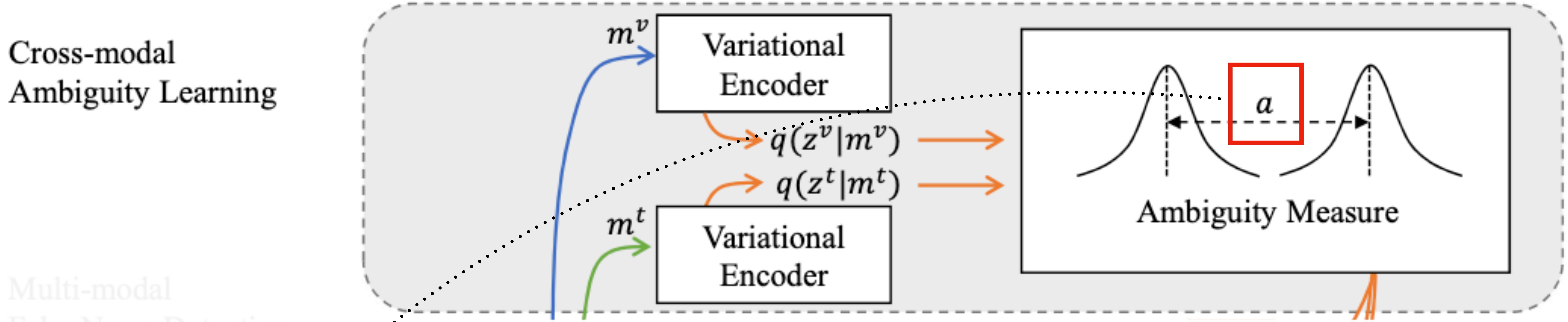
- Finally, jointly train the cross-modality alignment module to produce the semantically aligned unimodal representations m^t and m^v.
    - As input of cross-modal ambiguity learning module & cross-modal fusion module.



$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } y_2 = 0 \end{cases}$$

Methodology

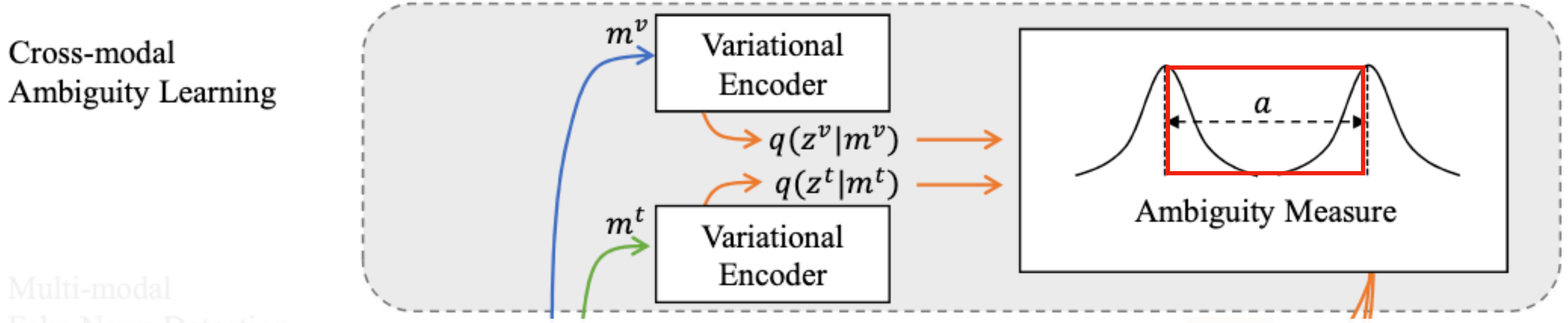
Cross-modal Ambiguity Learning



- Propose an ambiguity learning method via evaluating KL divergence between unimodal distributions approximated by two modal-specific variational auto-encoders (VAE).
- Learned ambiguity score is then used to adaptively control contribution of cross-modal & unimodal features in FND. Therefore, when unimodal features present strong ambiguity, cross-modal fake news detector should pay more attention to cross-modal features, and vice versa.

Methodology

Cross-modal Ambiguity Learning



- Assume the distributional divergence between unimodal features represent the information gap between uni-modalities.
  - i.e. Use the divergence over feature space to approximate their ambiguity.

Methodology

Cross-modal Ambiguity Learning

- For each data sample $\mathbf{x}_i$ with aligned textual feature and image feature, the variational posteriors of the two modalities can be defined as follows:

- $$q(z_i^t \mid m_i^t) = \mathcal{N}(z_i^t \mid \mu(m_i^t), \sigma(m_i^t))$$
  $$q(z_i^v \mid m_i^v) = \mathcal{N}(z_i^v \mid \mu(m_i^v), \sigma(m_i^v))$$

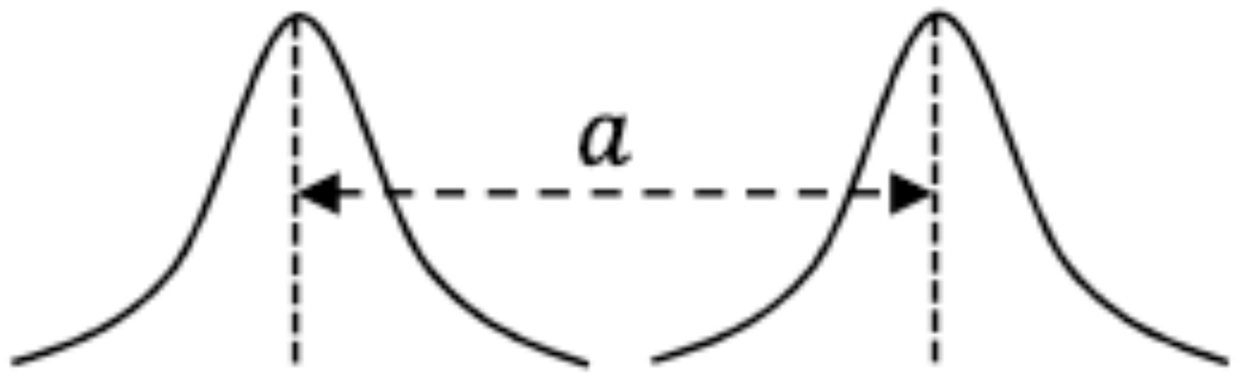
μ: mean, σ: variance

- Considering the distribution over the entire dataset, have

- $$q(z^t) = \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^t)}[q(z^t \mid m^t)] = \frac{1}{N}\sum_{i=1}^{N} q(z_i^t \mid m_i^t)$$
  $$q(z^v) = \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^v)}[q(z^v \mid m^v)] = \frac{1}{N}\sum_{i=1}^{N} q(z_i^v \mid m_i^v)$$

Methodology

Cross-modal Ambiguity Learning



- $D_{KL}(\cdot)$ denotes the KL divergence, and the ambiguity score a_i is computed as the symmetrized KL divergence obtained by averaging the normalized value.

- $$q(z_i^t \mid m_i^t) = \mathcal{N}(z_i^t \mid \mu(m_i^t), \sigma(m_i^t))$$
  $$q(z_i^v \mid m_i^v) = \mathcal{N}(z_i^v \mid \mu(m_i^v), \sigma(m_i^v))$$

  μ: mean, σ: variance

- $$q(z^t) = \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^t)}[q(z^t \mid m^t)] = \frac{1}{N}\sum_{i=1}^{N} q(z_i^t \mid m_i^t)$$
  $$q(z^v) = \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^v)}[q(z^v \mid m^v)] = \frac{1}{N}\sum_{i=1}^{N} q(z_i^v \mid m_i^v)$$

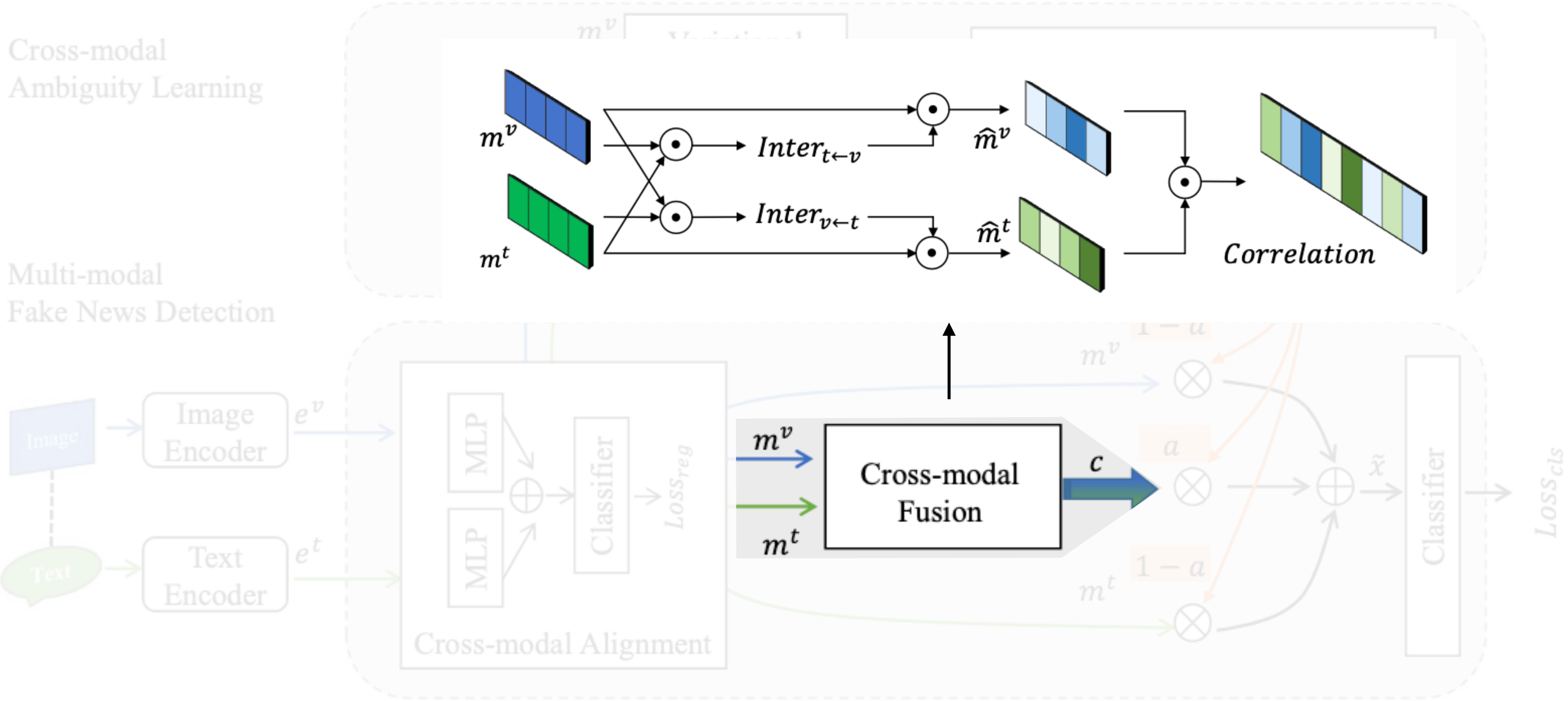
- $$a_i^1 = \frac{\mathrm{D}_{KL}(q(z_i^t \mid m_i^t) \| q(z_i^v \mid m_i^v))}{\mathrm{D}_{KL}(q(z^t) \| q(z^v))}$$
  $$a_i^2 = \frac{\mathrm{D}_{KL}(q(z_i^v \mid m_i^v) \| q(z_i^t \mid m_i^t))}{\mathrm{D}_{KL}(q(z^v) \| q(z^t))}$$
  $$a_i = \mathrm{sigmoid}(\frac{1}{2}(a_i^1 + a_i^2))$$

Methodology

Cross-modal Ambiguity Learning

- Small ambiguity score indicates that two unimodal distributions are close to each other.
- Utilize the ambiguity score a_i as the weight to govern the fusion of unimodal features and cross-modal features in both training and inference.
  - Cross-modal ambiguity learning help adaptively leverage cross-modal feature and drop out unimodal features when the ambiguity is large, and vice versa.

Methodology
Cross-modal Fusion

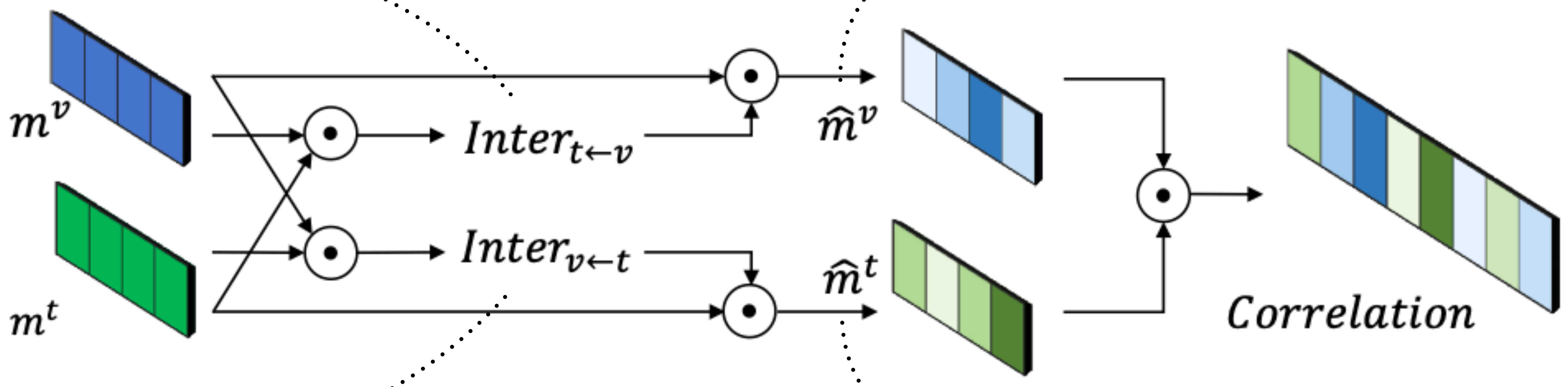


Methodology

Cross-modal Fusion

$$InterC_{t \leftarrow v} = \mathrm{softmax}([m^t][m^v]^T / \sqrt{\mathrm{dim}})$$

$$\hat{m}^v = InterC_{I \leftarrow T} \times m^v$$



$$c = \hat{m}^t \otimes \hat{m}^v$$

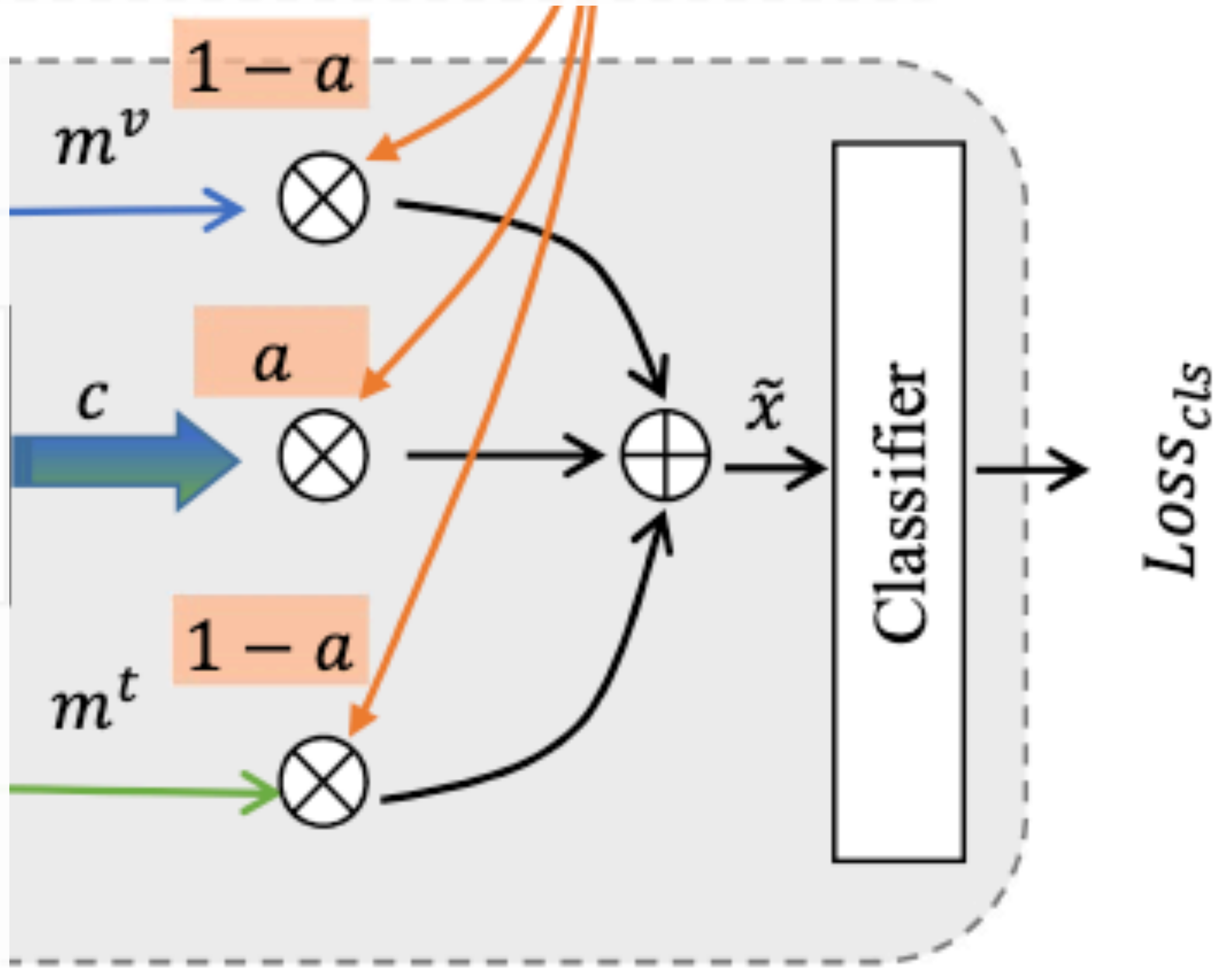
$$InterC_{v \leftarrow t} = \mathrm{softmax}([m^v][m^t]^T / \sqrt{\mathrm{dim}})$$

$$\hat{m}^t = InterC_{T \leftarrow I} \times m^t$$

Methodology

Classifier

- The input of the classifier is obtained by adaptively concatenating two sets of embeddings (m^t and m^v) governed by the cross-modal ambiguity score $a_{\mathbf{x}}$.
- Since fake news detection is a binary classification task, apply the cross-entropy loss.
- $\tilde{\mathbf{x}} = (a_{\mathbf{x}} \times c) \oplus ((1 - a_{\mathbf{x}}) \times m^t) \oplus ((1 - a_{\mathbf{x}}) \times m^v)$
- $\tilde{y}_1 = \mathrm{softmax}(MLP(\tilde{\mathbf{x}}))$
- $\mathscr{L}_{cls} = y_1 \log(\tilde{y}_1) + (1 - \tilde{y}_1)\log(1 - y_1)$



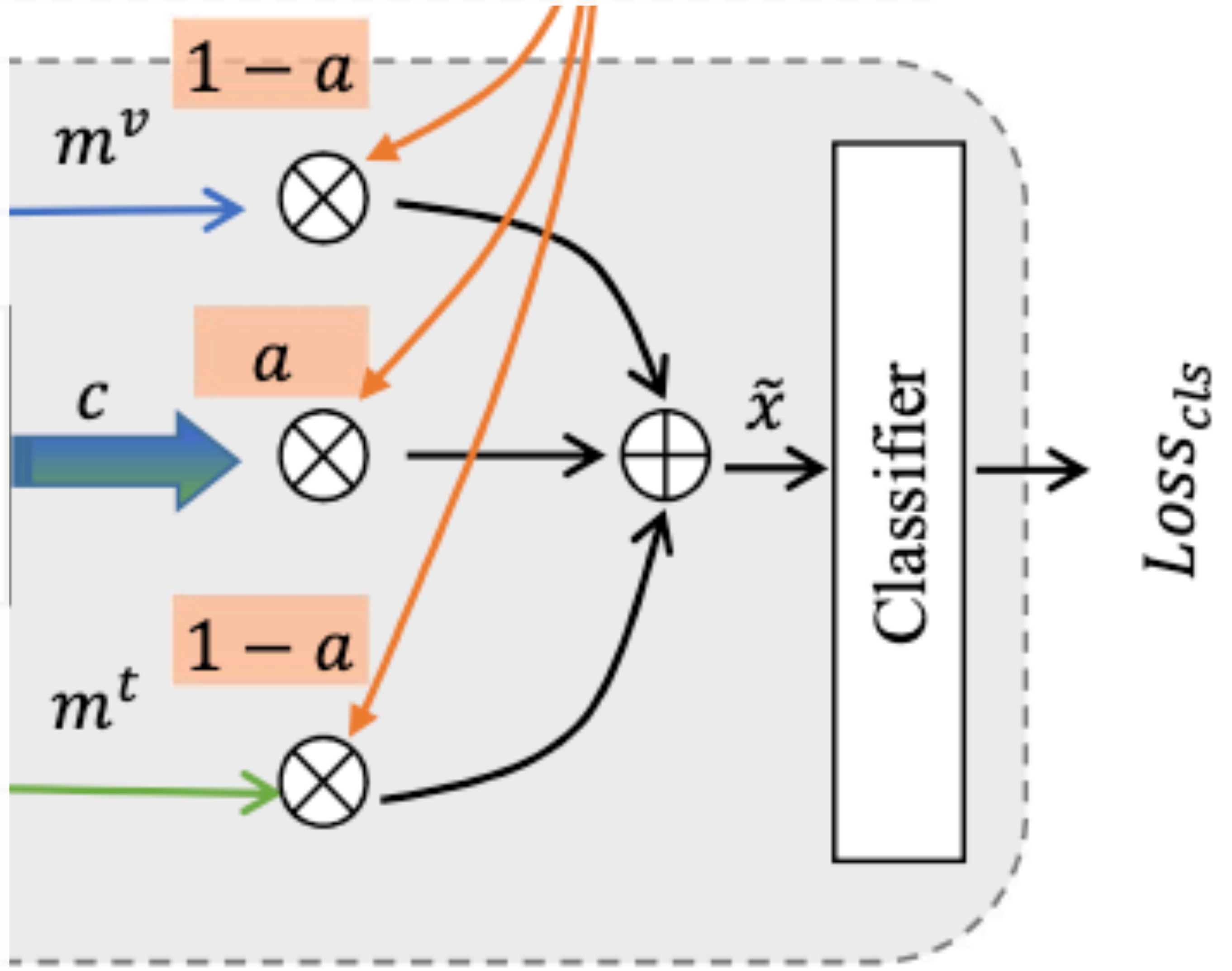
Methodology

Classifier

$$\mathcal{L}_{cls} = y_1 \log(\tilde{y}_1) + (1 - \tilde{y}_1)\log(1 - y_1)$$

$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos(e^t, e^v) & \text{if } \ y_2 = 1 \\ \max(0, \cos(e^t, e^v) - d) & \text{if } \ y_2 = 0 \end{cases}$$

- Next, discuss optimization strategy for the proposed method.
- Auxiliary semantic regularization task aims to bridge the semantic gaps between textual features and image features which may not be totally helpful for the classification task.
    - Limit its effect by placing a weight $\beta \in (0,1)$ on its loss function.
- Final loss function: $\mathcal{L} = \mathcal{L}_{cls} + \beta\mathcal{L}_{reg}$



Methodology

Algorithm of CAFE Training

Algorithm 1 Model training of CAFE.

Input: Datasets: $\mathcal{D}_1$ for the main task, $\mathcal{D}_2$ for the auxiliary task

Output: Model parameters: Θ_1 for the main task, Θ_2 for the auxiliary task

while not converge **do**
    for the auxiliary task **do**
        Sample minibatch from $\mathcal{D}_2$.
        Compute loss using $\mathcal{L}_{reg}\left(e^t, e^v, y_2\right)$.
        Update parameters in Θ_2 by Adam.
    end for
    for the main task **do**
        Sample minibatch from $\mathcal{D}_1$.
        Compute loss using $\mathcal{L}_{cls}\left(m^t, m^v, y_1\right)$.
        Update parameters in Θ_1 by Adam.
    end for
end while

Experiments

Datasets

- Twitter (MediaEval Verifying Multimedia Use task)
  - Training: 6840 real, 5007 fake
  - Test: 1406 posts
- Weibo
  - Training: 3783 real, 3749 fake
  - Test: 1996 posts

Experiments

Unimodal baselines

- CAR: combines RNN with attention mechanism to capture important textual features to detect text-only fake news.
- VS: explores visual and statistical features of image content to detect fake news.

Experiments
Multimodal baselines

- RA: LSTM network and attention mechanism to model text and social context.
- EANN: two related tasks: event discrimination and fake news detection.
- MVAE: variational autoencoder with a binary classifier to model representations.
- MKEMN: text, image and retrieved knowledge embeddings as stacked channels and makes a fusion via a convolutional operation.
- SAFE: pre-trained image to text model to transform image into text, then measures similarity.
- MCNN: textual semantic features, visual tampering features and similarity of textual and visual information computed by the cosine similarity.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- CAFE outperforms all the compared methods on every dataset in terms of Acc and F1.
- CAFE achieves the highest accuracy of 80.6% and 84.0% on two real-world datasets, respectively.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- Multimodal outperform the unimodal methods in all datasets.
- Confirming the advantage of leveraging multimodal information in fake news detection.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- RA and EANN perform worst.
- Because both methods learn uni-modality features separately and ignore the semantic gap across modalities resulting in different embedding spaces and less effective fusion.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- Performance of MKEMN varies significantly among different datasets.
- MKEMN regards different modalities as stacked channels without considering the heterogeneity issue, bonding its performance on the data distribution.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- MCNN achieves the best performance among all baselines.
- Due to the adoption of cross modality correlation captured by the cosine similarity between textual and visual features.

Experiments
Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- Auxiliary correlation learning task in CAFE can produce discriminative unimodal features.
- Ensure well aligned semantic space across different modalities and adaptively utilize these aligned features to assistant the main task.

Experiments

Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- Cross-modality ambiguity learning module can accurately estimate the ambiguity between different modalities.
- Weigh the importance between unimodal features and cross-modal features given different levels of ambiguity.

Experiments

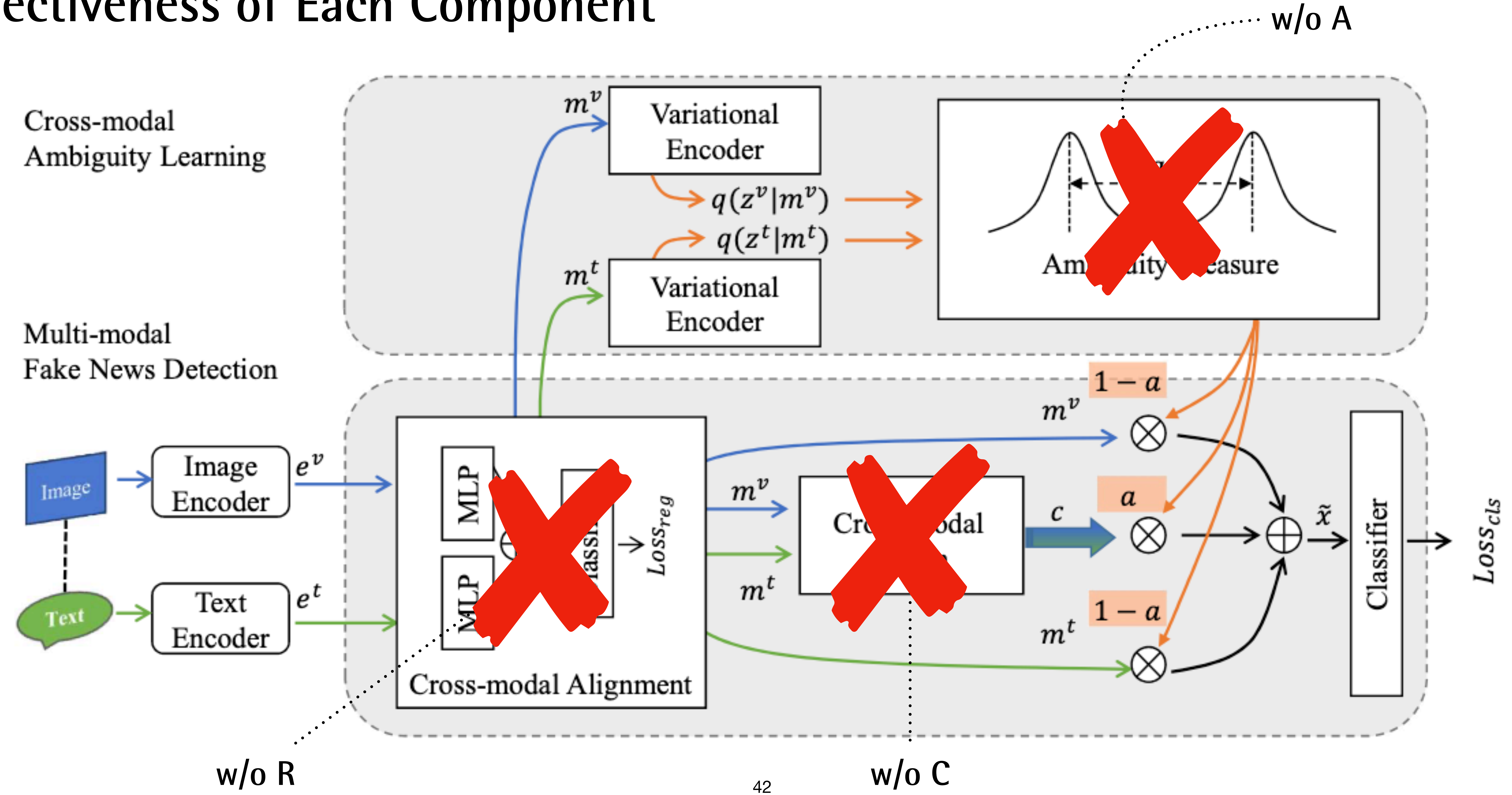
Overall Performance

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	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
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

- Main fake news detection task in CAFE can adaptively aggregate complementary unimodal representations and cross-modal correlations to perform accurate classification.

Experiments

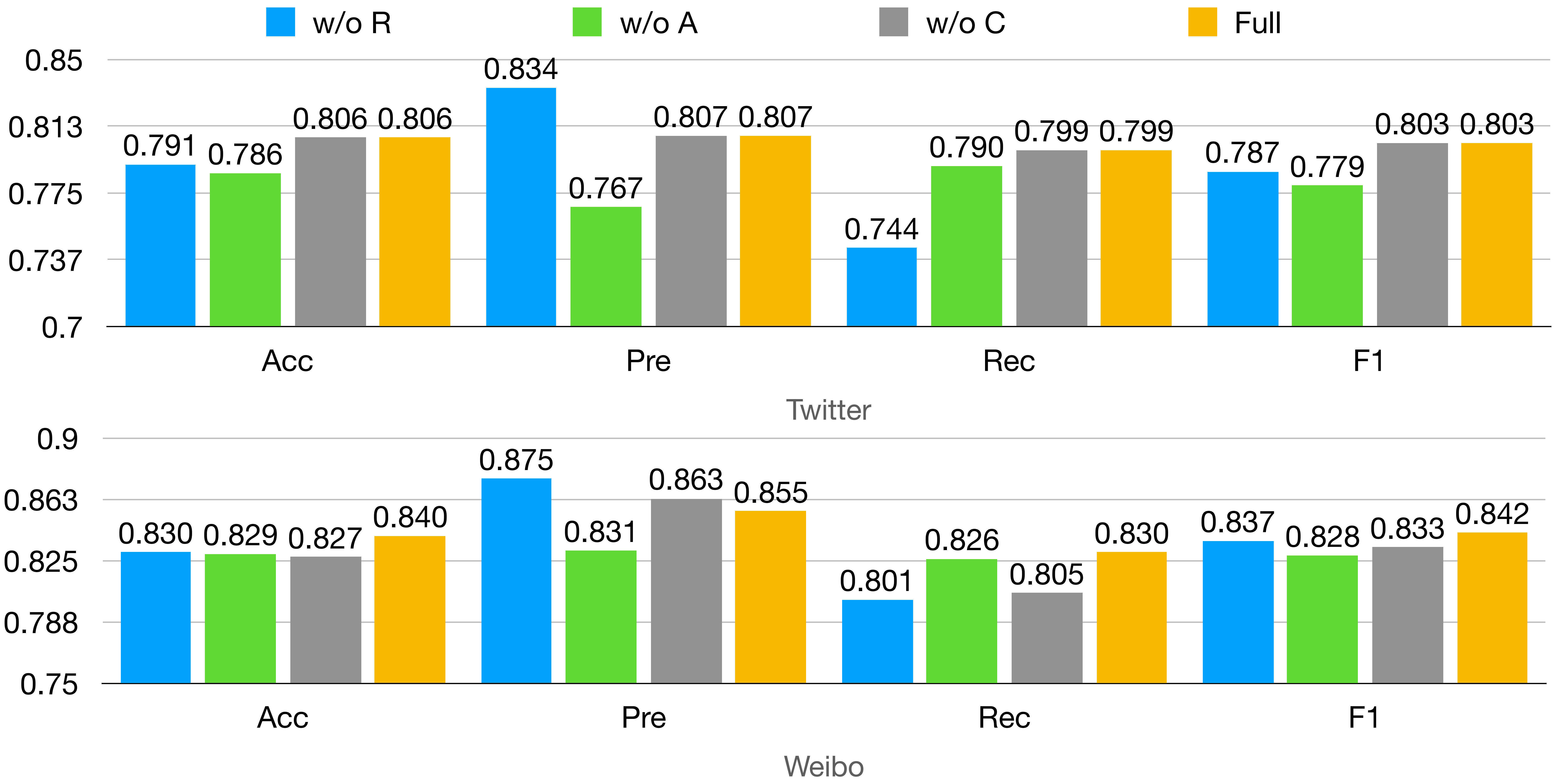
Effectiveness of Each Component



Experiments

Effectiveness of Each Component

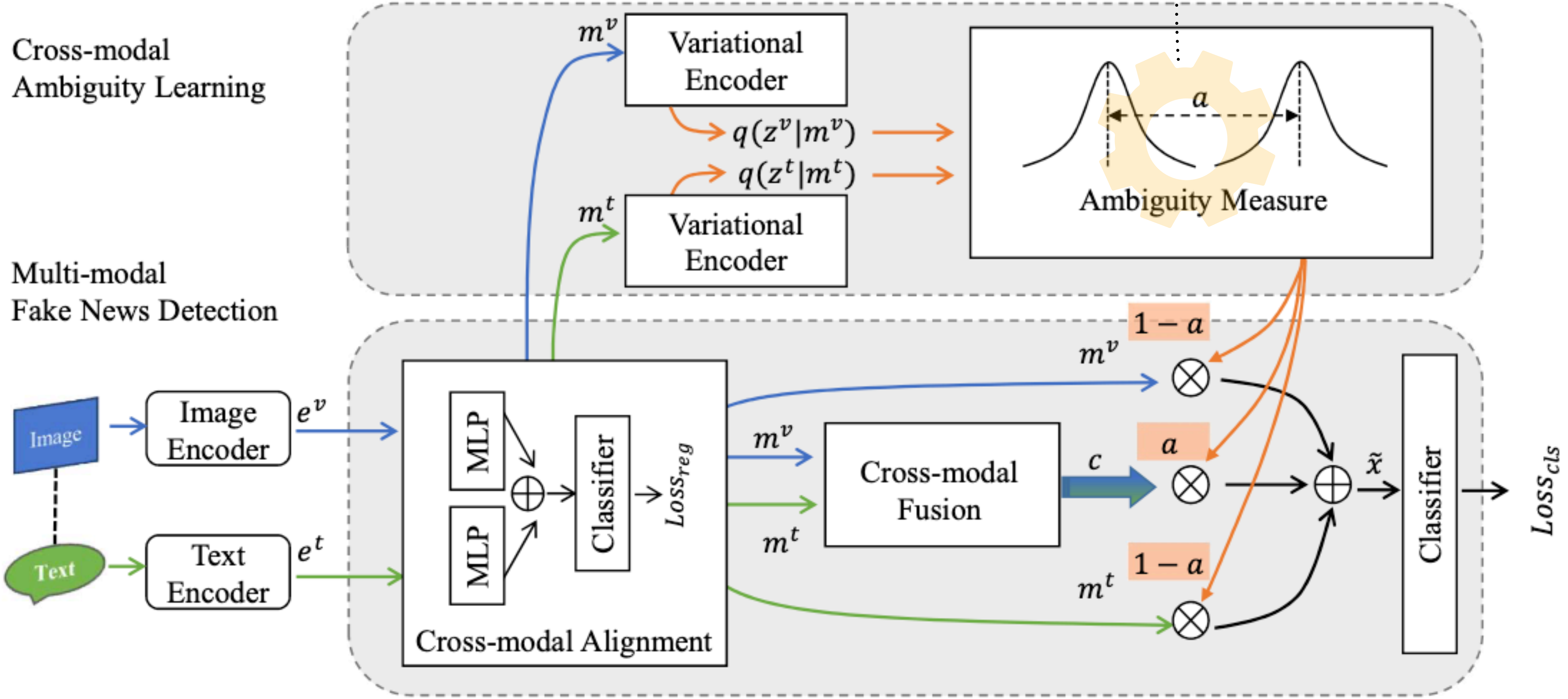
Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE w/o R	Twitter	0.791	0.834	0.744	0.787
	Weibo	0.830	0.875	0.801	0.837
CAFE w/o A	Twitter	0.786	0.767	0.790	0.779
	Weibo	0.829	0.831	0.826	0.828
CAFE w/o C	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.827	0.863	0.805	0.833
CAFE	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842



Experiments

Cross-modal Ambiguity Learning Analysis

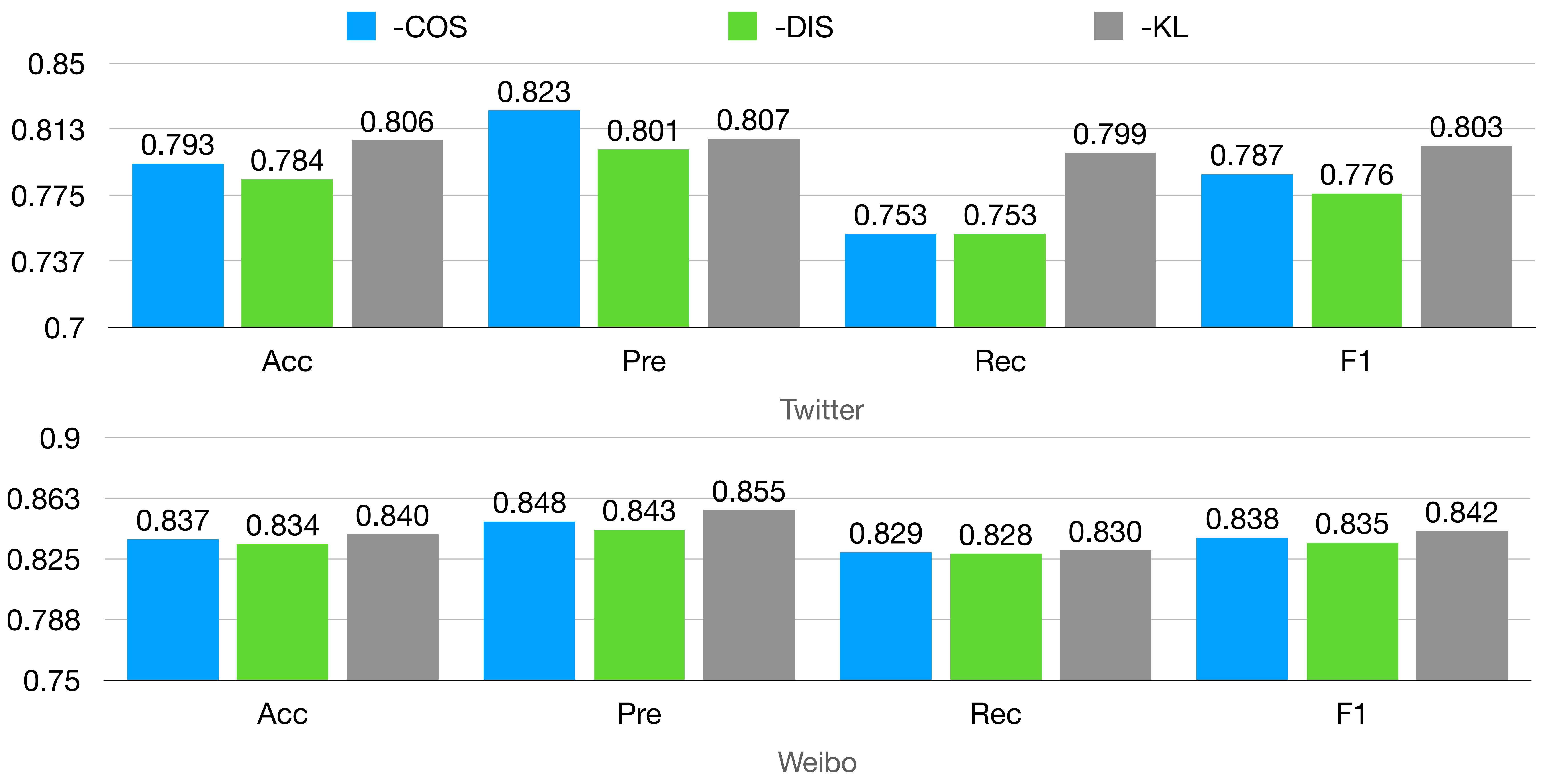
Change to:
COS - Cosine distance
DIS - Euclidean distance



Experiments

Cross-modal Ambiguity Learning Analysis

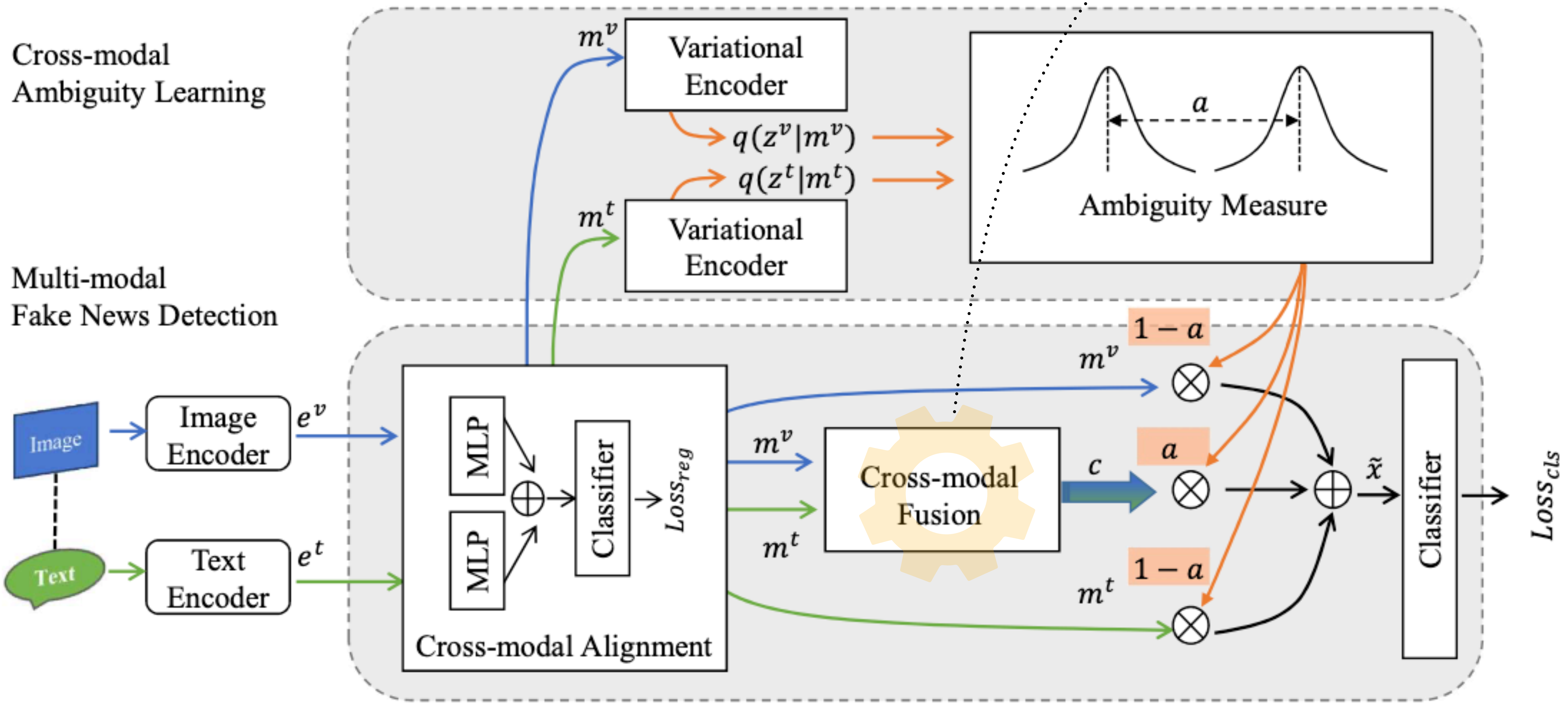
Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE-COS	Twitter	0.793	0.823	0.753	0.787
	Weibo	0.837	0.848	0.829	0.838
CAFE-DIS	Twitter	0.784	0.801	0.753	0.776
	Weibo	0.834	0.843	0.828	0.835
CAFE-KL	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842



Experiments

Cross-modal Fusion

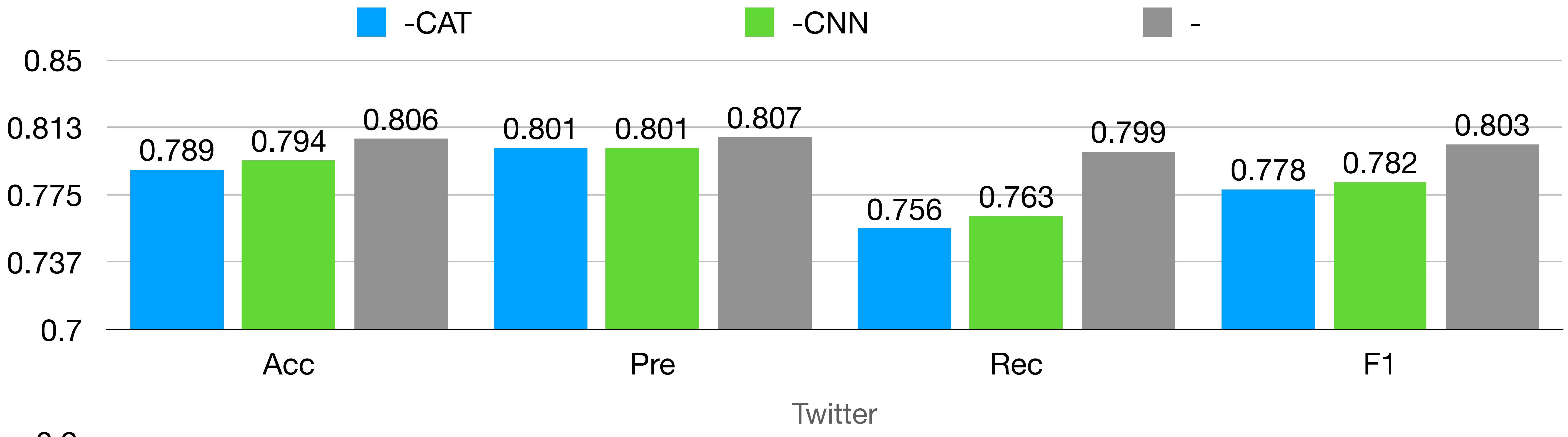
Change to:
CAT - Concatenate
CNN

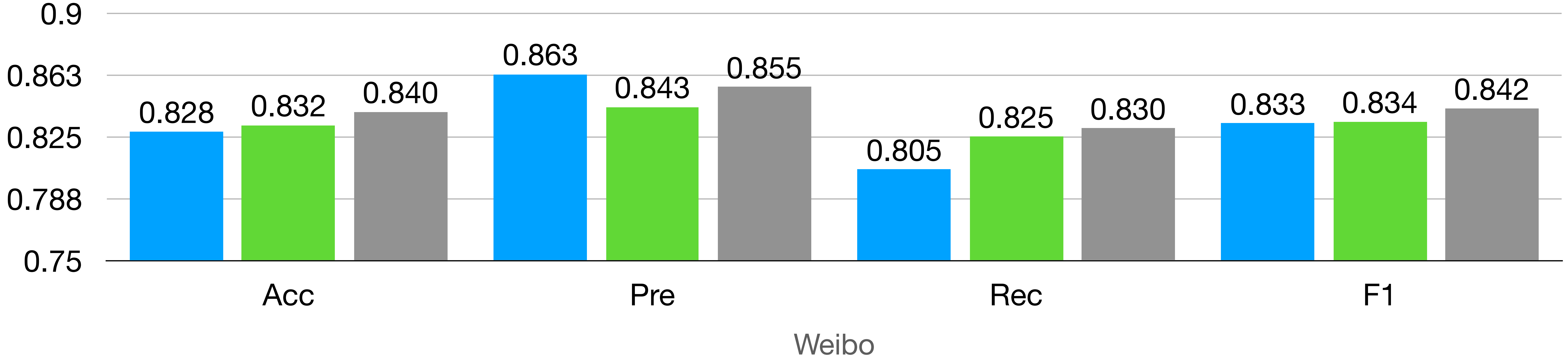


Experiments

Cross-modal Ambiguity Learning Analysis

Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE-CAT	Twitter	0.789	0.801	0.756	0.778
	Weibo	0.828	0.863	0.805	0.833
CAFE-CNN	Twitter	0.794	0.801	0.763	0.782
	Weibo	0.832	0.843	0.825	0.834
CAFE	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842





Experiments

Quantitative analysis

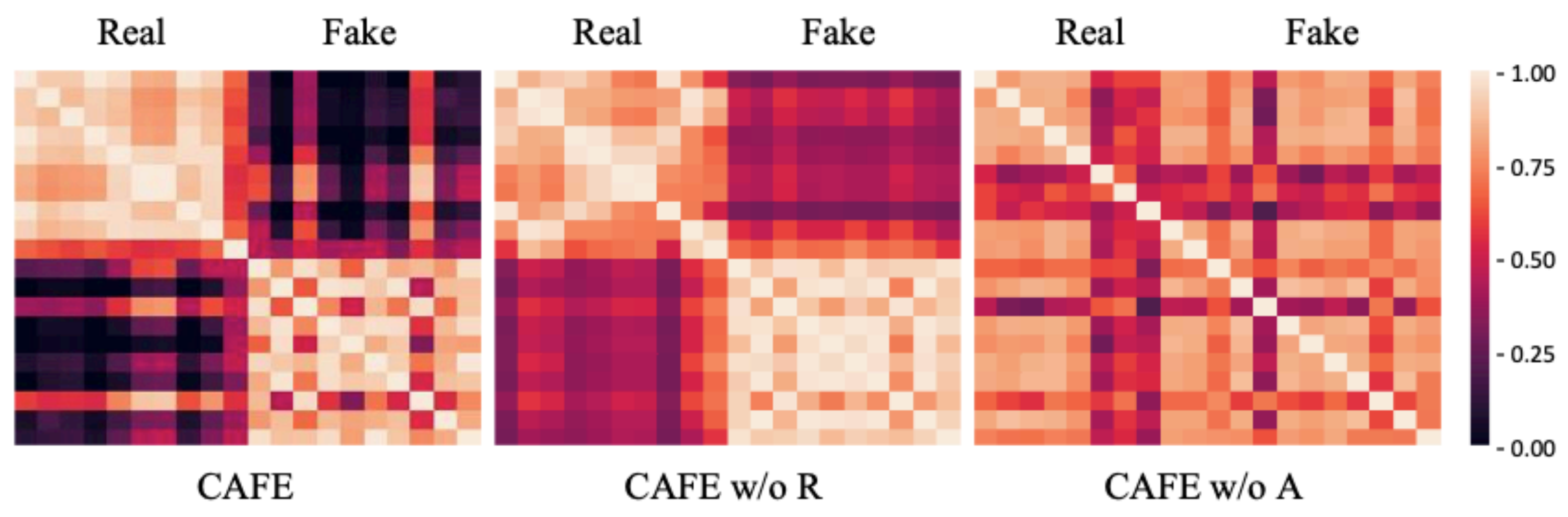


Figure 5: The result of quantitative analysis. CAFE presents clear inter-class difference and intra-class similarity, while CAFE w/o A and CAFE w/o R yield poor capability to learn inter-class difference.

- Use heat-maps to visualize the correlation patterns between inter-class and intra-class news.
- Select 20 news, including 10 pieces of fake news and 10 pieces of real news, and then extract the corresponding correlations from CAFE, CAFE w/o R & CAFE w/o A.
- CAFE can learn the discriminative cross-modal features which are explicitly beneficial to the cases when uni-modalities present strong ambiguity, and thus improve multimodal fake news detection accuracy.

Conclusion
of CAFE

- Cross-modal ambiguity is crucial in multimodal fake news detection.
    - Formulate the cross-modal ambiguity learning task.
- Proposed CAFE which is capable of adaptively aggregating discriminative cross-modal correlation features and unimodal features based on the inherent cross-modal ambiguity.
    - Addressing the misclassifications caused by the disagreement between different modalities.
- Experimental on two datasets demonstrate that CAFE outperforms in multimodal FND.

Comments
of CAFE

- Propose concept of cross-modal ambiguity.
- May can re-design the calculation of ambiguity to improve the performance.

→ How to integrate w/ knowledge graph?

→ can knowledge graph can filter some redundant information?