



Multi-Modal Co-Training for Fake News Identification Using Attention-Aware Fusion

Sreyasee Das Bhattacharjee and Junsong Yuan

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

November 10, 2021

Multimodal Co-training for Fake News Identification using Attention-aware Fusion

Sreyasee Das Bhattacharjee and Junsong Yuan

State University of New York at Buffalo, USA

Abstract. Rapid dissemination of fake news to purportedly mislead the large population of online information sharing platforms is a societal problem receiving increasing attention. A critical challenge in this scenario is that a multimodal information content, e.g., supporting text with photos, shared online, is frequently created with an aim to attract attention of the readers. While ‘fakeness’ does not exclusively synonymize ‘falsity’ in general, the objective behind creating such content may vary widely. It may be for depicting additional information to clarify. However, very frequently it may also be for propagating fabricated or biased information to purposefully mislead, or for intentionally manipulating the image to fool the audience. Therefore, our objective in this work is evaluating the veracity of a news content by addressing a two-fold task: (1) if the image or the text component of the content is fabricated and (2) if there are inconsistencies between image and text component of the content, which may prove the image to be out of context. We propose an effective attention-aware joint representation learning framework that learns the comprehensive fine-grained data pattern by correlating each word in the text component to each potential object region in the image component. By designing a novel multimodal co-training mechanism leveraging the class label information within a contrastive loss-based optimization, the proposed method exhibits a significant promise in identifying cross-modal inconsistencies. The consistent out-performances over other state-of-the-art works (both in terms of accuracy and F1-score) in two large-scale datasets, which cover different types of fake news characteristics (defining the information veracity at various layers of details like ‘false’, ‘false connection’, ‘misleading’, and ‘manipulative’ contents), topics, and domains demonstrate the feasibility of our approach.

Keywords: Fake News Detection, Rumor, Multimodal Classification, Co-training, Attention, Feature Fusion

1 Introduction

The task of Fake news detection is to identify deceptive digital news content in the web-based platforms. With an abundance of information available from competing resources, it is often difficult for users to gauge the veracity of an

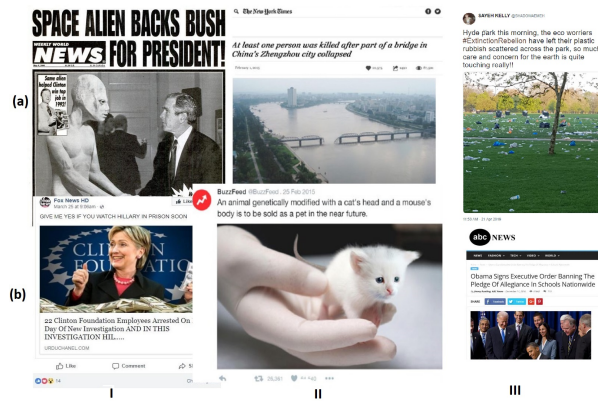


Fig. 1. Examples of some real instances of Fake News articles which use Multimodal content to dupe readers. In (a)I, (b)I, (a)II, and (b)II, images were purposely manipulated to describe a fake article. The instances in (a)III and (b)III represent some out of context images, so the images do not accurately support their text descriptions.

online news content in a timely manner. While, per Gallup poll¹, only 40% of the Americans trust their mass media resources to report the news ‘fully, accurately and fairly’, a critical bias towards Internet based resources (like blogs and social media) still prevails. Such alternative digital information resources leave the readers more susceptible to incomplete and deceptive information [1].

Figure 1 shows some instances ((a)I, (a)II, (b)I, and (b)II) of fake news, where image components were purportedly forged/morphed to support fake news contents. Unlike these standard methods for misrepresentations, to obfuscate the usual fact checking software, recent trends have been using out of context images as shown in the right column images ((a)III, (b)III), where a manipulative correlation is generated between two components to propagate disinformation. So for this type of fake-news contents, relevance of the text topic of the news content with its image component is not pertinent. Therefore, our objective in this work is evaluating the veracity of a news content by addressing a two-fold task: (1) if the image or the text component of the content is fabricated and (2) if there are inconsistencies between image and text component of the content, which may prove the image to be out of context.

A good set of works have leveraged traditional machine learning methods as well as recent deep learning models [10, 23, 24, 29], most of which rely on textual content or other metadata (like news source, emotional features, number of likes, etc.) including content creator’s profile information. While a few recent works have proposed multimodal methods to address the task of fake news detection [11, 13, 15, 30], majority of these methods just combine different mode-specific feature vectors (visual, textual features, and semantic information), derived from some independently customized pre-trained models. Therefore, the

¹ <https://news.gallup.com/poll/321116/americans-remain-distrustful-mass-media.aspx>

mutual relation between these mode-specific representations and how they may jointly describe the veracity trait of a news content, are still under-explored.

We argue that grounding the text component to different semantic areas in the image component of a multimodal news content is crucial for evaluating the veracity of its information. Additionally, to gauge the quality of the visual component, interaction between different object regions and relative position information within its visual component often play a critical role. Finally, in order to ‘bridge the gap’ between the complementary modes, it is important to identify the hard positives, which may be leveraged to enhance the contrastive characteristics of the learned joint representation, so the clusters of points belonging to the same class are pulled together in the learned embedding space, while simultaneously pushing apart clusters of samples from different classes. To this end, the contributions of this work include the followings:

- 1.To ensure a more accurate evaluation of the quality of the visual and text component of a news content, the proposed method learns the localized data patterns by leveraging self- and cross-modal attentions at multiple layers of details. This enables capturing the correlation of each word to each potential object region within the image component, within the learned initial joint representation, while ascertaining an enhanced decision interpretability in parallel.
- 2.To analyze the cross-modal inconsistencies, the initial joint representation is further finetuned by a multimodal co-training scheme that leverages the label information within a formulation of the supervised contrastive learning framework, to explicitly capture the complementary category relevant information and their mutual interaction observed within the different mode representations of the same data source.
- 3.The proposed method is extensively evaluated on two benchmark datasets (Twitter Dataset, as part of MediaEval [4], which was released for evaluating methods for detecting fake multimedia content in social media and the large scale Fakeddit Dataset [21] that has samples from up to six different categories of information disorder) and it consistently outperforms other state-of-the-art approaches both in terms of accuracy and F1-score.

The rest of the paper is organized as follows: Section 2 briefly describes related works. The proposed method is explained in Section 3. Section 4 and 5 respectively present the experimental results and conclusion.

2 Related Works

The existing set of literature addressing the task of Fake news recognition may be split into two groups: uni-mode methods [3, 20] and multimodal methods [15, 30, 32]. Many early works utilizing only the text-based features, retrieve linguistic features (like special characters, emojis) [3, 20] or language stylistic features (like assertive verbs, discourse markers) [22] to assess the credibility of the news content. Researchers have also explored the role of profile information [9, 24], emotions [8], social context features [13], and source credibility [6] to

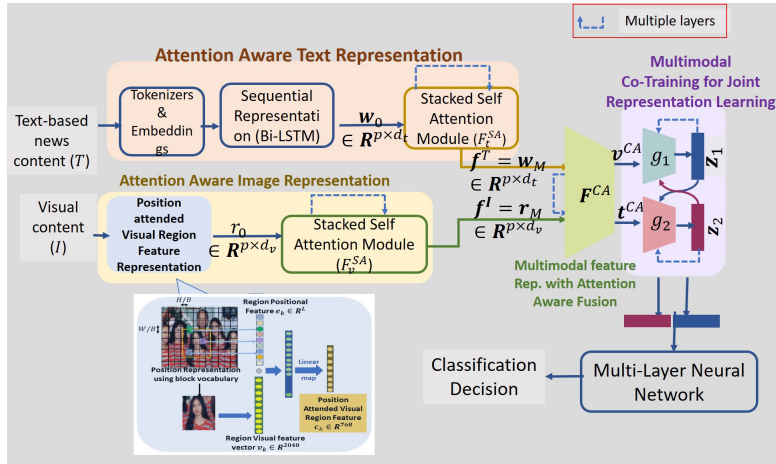


Fig. 2. The Proposed Method Overview is shown in (a) and the *Position attended Visual Region Feature Representation* scheme is illustrated in (b).

evaluate the factual quality of the post. In a recent work, Lago et al. [19] have evaluated different methods for identifying manipulated images.

Recent studies have shown that analyzing the accompanying visual component may improve the fake-news detection task [13, 21, 26, 30]. Singhal et al. [26] propose a multimodal fake news detection framework that employs a language transformer and visual module warmed up by the pre-trained CNNs to derive mode-specific features, which are later concatenated to obtain a learned multimodal feature descriptor. Wang et al. [30] propose an adversarial network utilizing a two-stream event invariant feature extractors (text-based CNN for text mode and VGG-19 based network for visual mode). Both descriptors learned independently using the samples' mode-specific representations are just concatenated for designing a single feature vector. Similarly, Nakamura et al. [21] utilize a two-stream network for processing textual (using bidirectional BERT model) and visual (using ResNet50 model) information. Authors analyze three fixed feature fusion techniques (maximum, average, and concatenate) for their performance in fake news detection task. Another set of works also leverage textual, visual and metadata [5, 13, 18]. However, the relation between these mode-specific representations and how they may jointly interact to evaluate the veracity trait of a news content are still not sufficiently investigated, which form the foundation of our method.

3 Proposed Method

Figure 2 presents the entire workflow of the proposed multimodal Fake-news recognition system. In this paper, the proposed method is described for two-mode

(image, text) data representations. However, the learning strategy is completely generic, such that the extension to a higher number of modes is straightforward.

In the following representation, the annotated data collection is expressed as: $\mathcal{D} = \{(\mathbf{s}^j, l^j)\}_j$, where a sample \mathbf{s}^j is represented in terms of an image-text pair (I^j, T^j) , such that I^j and T^j respectively represent the image and the text component \mathbf{s}^j from category $l^j \in \mathcal{C}$. The objective of this work is to learn an effective classifier model that can evaluate the veracity of the news content passed as a query during test time.

3.1 Attention Aware Image Representation

While usually for the representation task, the entire image is considered as a single quantity, which is processed using a CNN [12], for a richer and more detailed understanding of the image, in this work, we aim to divide the image into semantically meaningful regions to derive a region-level representation.

Region Visual Features: Given a multimedia data content (I, T) as a sample, its image component I is represented in terms of n candidate interest regions $\{v_k\}_{k=1}^n$ called ‘proposals’, generated by Selective Search [27]. Each region depicts a potential object region within the image I and is represented in terms of a CNN vector, i.e. $I = \{\mathbf{v}_k\}_{k=1}^n$, where $\mathbf{v}_k \in \mathbb{R}^{2048}$ is the CNN feature vector representing the k^{th} region v_k within I . In our experiments, image region representation was extracted from VGG-19 model pre-trained on ImageNet [25], which was dimension reduced using Principal Component Analysis (PCA) to a 2048 dimension feature [14].

Region Positional Features: Intuitively, the position information of the image regions and their relative placements in the whole image are very important. Motivated by this thought, we combine the position information within the learning procedure to capture more accurate and fine-grained image spatial layout information. For any given multimedia data content (I, T) , its image component I is first resized to a pre-defined size $W \times H$. Then we adopt the position representation approach by Wang et al. [31] to equally split into $B \times B$ blocks. This collection of split blocks forms the position vocabulary to represent the positional information of every region within I . The position of each split block represented by the one-hot vector of dimension B^2 , indicating its index in the position vocabulary. Finally, the L dimensional region position vector for each v_k defined by an embedding vector \mathbf{e}_k , is defined as $\mathbf{e}_k[l] = OV(b_l, v_k) \times f(b_l, v_k)$ where $f(b_l, v_k)$ computes the spatial proximity between the region v_k and the l^{th} max overlapping block b_l ($1 \leq l \leq L$) and $OV(b_l, v_k)$ computes their normalized pairwise overlap ratio. In order to capture the spatial information of each region and preserve the locality within the derived embedding vector (i.e. the neighboring position embeddings should be similar), we encode the normalized position information of v_k as: $\mathbf{p}_k = [N(x_k), N(y_k), N(w_k), N(h_k)]$, where (x_k, y_k, w_k, h_k) denotes the (x-location, y-location, width, height) of v_k and $N(\cdot)$ normalizes the input within a range of 0 and 1. The spatial proximity between b_l and v_k is computed as $f(b_l, v_k) = e^{-\frac{\|\mathbf{p}_k - \mathbf{p}_l\|}{2\sigma^2}}$, where \mathbf{p}_l represents the normalized position information of the position vocabulary box b_l and $\sigma \in \mathbb{R}$ is a scalar, set as the

average distance between all the position embeddings. The region position vector \mathbf{e}_k is concatenated with the region visual feature \mathbf{v}_k and pass through a linear layer to obtain the final d_v dimensional *Position attended Visual Region Feature* vector $\mathbf{c}_k \in \mathbb{R}^{d_v}$ ($d_v = 768$). The process is illustrated in Figure 2.

Image Representation: The collection of $\{\mathbf{c}_k\}_{k=1}^n$ describes potential object regions within the image I and passed through a feature representation module of F_v^{SA} , comprising of a stack of M self attention layer based units, to derive an attention aware final image representation vector \mathbf{f}^I . More precisely, given the initial region-based representation of I as $\mathbf{r}_0 = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_n]$, the m^{th} -layer processing of F_v^{SA} , which takes $\mathbf{r}_0 \in \mathbb{R}^{n \times d_v}$ as input at the layer $m = 0$, is performed as $\mathbf{r}_{m+1} = \text{linear}\left(\text{Softmax}\left(\frac{\mathbf{r}_m \mathbf{r}_m^T}{\sqrt{d_v}}\right) \mathbf{r}_m\right)$. The final M^{th} layer output $\mathbf{r}_M = F_v^{SA}(\mathbf{r}_0, M) \in \mathbb{R}^{n \times d_v}$, is used to represent the image component I and we set $\mathbf{f}^I = \mathbf{r}_M$.

3.2 Attention Aware Text Representation

The proposed textual embedding processing module maps each word to a high dimensional vector space. Each input text component T is considered as a sequence of p words, i.e. $T = \{\omega_1, \omega_2, \dots, \omega_p\}$. We employ the pretrained model BERT [7] to obtain the fixed word embedding vector of size d_w . To capture the contextual information of each word, we employ Bi-LSTM following the embedding layer, which has a forward hidden state $\overrightarrow{\mathbf{h}}_w \in \mathbb{R}^{d_w^{hid}}$ and the backward hidden state $\overleftarrow{\mathbf{h}}_w \in \mathbb{R}^{d_w^{hid}}$, where d_w^{hid} is the number of hidden units. For each ω_i , we concatenate both its forward and the backward hidden state representation to derive the final word representation vector $\mathbf{h}_i = [\overrightarrow{\mathbf{h}}_w, \overleftarrow{\mathbf{h}}_w] \in \mathbb{R}^{2d_w^{hid}}$ ($d_w^{hid} = 384$). Therefore, the text component T is represented in terms of a sequence of word descriptors $\mathbf{w}_0 = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_p]$, where $\mathbf{h}_i \in \mathbb{R}^{d_t}$ with $d_t = 2d_w^{hid} = 768$. Similar to the image representation technique, the proposed attention aware text representation module of F_t^{SA} is also comprised of a stack of self attention layer based units followed by a linear layer, which takes \mathbf{w}_0 as input at layer $m = 0$ to learn an attention aware text representative. The output of the m^{th} layer in F_t^{SA} is taken as an input to its $(m + 1)^{th}$ layer, which is computed as $\mathbf{w}_{m+1} = \text{linear}\left(\text{Softmax}\left(\frac{\mathbf{w}_m \mathbf{w}_m^T}{\sqrt{d_t}}\right) \mathbf{w}_m\right)$. The output of the last layer (we have considered a stack of M self attention layers for both visual and text representations), computed as $\mathbf{w}_M = F_t^{SA}(\mathbf{w}_0, M) \in \mathbb{R}^{p \times d_t}$, is used as the attention aware text representative for T and we set $\mathbf{f}^{T^{xt}} = \mathbf{w}_M$.

3.3 Multimodal Feature Representation with Attention Aware Fusion

For the fake news recognition task, the content is primarily described using two-mode representation: visual and text. Therefore, to obtain an effective multimodal representation of the content, it is important to integrate two mode-specific representations in such way that would preserve a detailed understanding of its text component and the relevance of the visual information accompanying the text. Therefore, directly concatenating the mode-specific features [21]

may not be the most effective information fusion approach for this application setting. In this work, we propose a multimodal fusion method that leverages cross-modal attention information to fuse these two mode-specific features \mathbf{f}^I and $\mathbf{f}^{T^{xt}}$. The proposed fusion module F^{CA} , consists of a stack of M multimodal attention layers (i.e. the key and value pair represents one mode and the query represents another mode) which merges two parallel and independent attention submodules to jointly learn two sophisticated multimodal attention enhanced visual and text representations denoted by \mathbf{v}_M^{text} and \mathbf{t}_M^{vis} respectively. At the 0^{th} layer of F^{CA} , \mathbf{f}^I and $\mathbf{f}^{T^{xt}}$ respectively represent visual and text-based representation of the given multimodal sample (I, T) i.e., $\mathbf{v}_0^{text} = \mathbf{f}^I \in \mathbb{R}^{n \times d_v}$ and $\mathbf{t}_0^{vis} = \mathbf{f}^{T^{xt}} \in \mathbb{R}^{p \times d_t}$. Also, the entire M -layer non-linear multimodal fusion module can be represented as $(\mathbf{v}_M^{text}, \mathbf{t}_M^{vis}) = F^{CA}(\mathbf{f}^I, \mathbf{f}^{T^{xt}})$. At any intermediate m^{th} layer of F^{CA} , the attention enhanced joint embeddings are simultaneously updated as:

$$\mathbf{v}_{m+1}^{text} = linear \left(Softmax \left(\frac{\mathbf{V}\mathbf{T}_m(\mathbf{V}\mathbf{T}_m)^T}{\sqrt{d_v}} \right) \mathbf{V}\mathbf{T}_m \right) \quad (1)$$

$$\mathbf{t}_{m+1}^{vis} = linear \left(Softmax \left(\frac{\mathbf{T}\mathbf{V}_m(\mathbf{T}\mathbf{V}_m)^T}{\sqrt{d_t}} \right) \mathbf{T}\mathbf{V}_m \right) \quad (2)$$

where we have $\mathbf{V}\mathbf{T}_m = linear(Softmax(\frac{\mathbf{v}_m^{text}(\mathbf{t}_m^{vis})^T}{\sqrt{d_v}})\mathbf{t}_m^{vis})$ and similarly compute $\mathbf{T}\mathbf{V}_m = linear(Softmax(\frac{\mathbf{t}_m^{vis}(\mathbf{v}_m^{text})^T}{\sqrt{d_t}})\mathbf{v}_m^{text})$. In fact, for any intermediate m^{th} layer, the $\mathbf{V}\mathbf{T}_m$ (and $\mathbf{T}\mathbf{V}_m$) is the multimodal scaled dot product attention that derives the semantic (and visual) context for a *multimodal attention enhanced visual* (and *text*) feature $\mathbf{v}_{m+1}^{text} \in \mathbb{R}^{n \times d_t}$ (and $\mathbf{t}_{m+1}^{vis} \in \mathbb{R}^{p \times d_v}$).

Finally, both \mathbf{v}_M^{text} and \mathbf{t}_M^{vis} are average pooled along their respective first dimension, to obtain an aggregated *multimodal attention enhanced visual* representation $\mathbf{v}^{CA} = MeanPool(\mathbf{v}_M^{text}) \in \mathbb{R}^{d_t}$ and an aggregated *multimodal attention enhanced text* representation $\mathbf{t}^{CA} = MeanPool(\mathbf{t}_M^{vis}) \in \mathbb{R}^{d_v}$. The function $MeanPool(\cdot)$ is the average pooling function along the first dimension of its argument.

3.4 Multimodal Co-Training for Joint Representation Learning

Each sample $(\mathbf{s}^j, l^j) \in \mathcal{D}$ is now represented using two views, $\mathbf{s}^j = \{\mathbf{v}^{CA,j}, \mathbf{t}^{CA,j}\}$, where $\mathbf{v}^{CA,j}$ refers to the *multimodal attention enhanced visual* feature space representation of I^j (the image component of \mathbf{s}^j) and $\mathbf{t}^{CA,j}$ refers to the *multimodal attention enhanced text* feature space representation of T^j (the text component of \mathbf{s}^j). To further highlight the cross-modal consistency utilizing the label information, the objective of multimodal co-training is to learn two functions g_1 and g_2 , where $\mathbf{z}_1^j = g_1(\mathbf{v}^{CA,j})$ and $\mathbf{z}_2^j = g_2(\mathbf{t}^{CA,j})$, so that both \mathbf{z}_1^j and \mathbf{z}_2^j would emit higher similarity scores with the elements of a hard positive set \mathcal{P}^j while simultaneously enhancing their differences with the elements of the negative set \mathcal{N}^j .

We propose to co-train these models (g_1 and g_2) by retrieving the hard positives from the other view representation. The function g_1 is updated with a su-

pervised contrastive loss [16] computed using a random batch of samples $\mathcal{B} \subset \mathcal{D}$:

$$\mathcal{L}_1 = - \sum_{i \in \mathcal{B}} \mathbb{E} \left[\log \frac{\sum_{p \in \mathcal{P}_1^j} \exp(\mathbf{z}_1^j \cdot \mathbf{z}_1^p / \tau)}{\sum_{p \in \mathcal{P}_1^j} \exp(\mathbf{z}_1^j \cdot \mathbf{z}_1^p / \tau) + \sum_{n \in \mathcal{N}_1^j} \exp(\mathbf{z}_1^j \cdot \mathbf{z}_1^n / \tau)} \right] \quad (3)$$

where the numerator is defined as a sum of ‘similarity’ between \mathbf{z}_1^j (i.e. the g_1 transformed output of $\mathbf{v}^{CA,j}$) and a positive set \mathcal{P}_1^j , constructed by identifying the most similar samples using its corresponding *multimodal attention enhanced text* feature $\mathbf{t}^{CA,j}$. The term $\tau \in \mathbb{R}^+$ is scalar temperature parameter. By leveraging the label information in a supervised scenario, the structure of \mathcal{P}_1^j is defined as follows:

$$\mathcal{P}_1^j = \{A^{vis}(I^j, a), I^k | k \in Nbr_K(\mathbf{z}_2^j \cdot \mathbf{z}_2^i), \forall (\mathbf{s}^i, l^i) \in \mathcal{D} \setminus \{(\mathbf{s}^j, l^j)\}, l^k == l^j, a \in \mathcal{A}^{vis}\} \quad (4)$$

where $A^{vis}(I^j, a)$ obtains the augmented version of I^j , parameterized by a sampled from a pre-defined set of augmentation transformations in \mathcal{A}^{vis} , $Nbr_K(x, \cdot)$ identifies the indices of top K similar samples to x and $\mathbf{z}_2^j \cdot \mathbf{z}_2^i$ computes the similarity between \mathbf{z}_2^j and \mathbf{z}_2^i . Hence the \mathcal{P}_1^j consists of top- K similar samples to \mathbf{s}^j retrieved using their g_2 transformed *multimodal attention enhanced text* feature space representation plus the j^{th} sample’s own augmentations, and \mathcal{N}_1^j represents the complement of \mathcal{P}_1^j that does not include samples with the same annotation l^j and their augmentations.

Similarly the function g_2 is updated with a similar supervised contrastive loss computed using a random batch of samples $\mathcal{B} \subset \mathcal{D}$:

$$\mathcal{L}_2 = - \sum_{i \in \mathcal{B}} \mathbb{E} \left[\log \frac{\sum_{p \in \mathcal{P}_2^j} \exp(\mathbf{z}_2^j \cdot \mathbf{z}_2^p / \tau)}{\sum_{p \in \mathcal{P}_2^j} \exp(\mathbf{z}_2^j \cdot \mathbf{z}_2^p / \tau) + \sum_{n \in \mathcal{N}_2^j} \exp(\mathbf{z}_2^j \cdot \mathbf{z}_2^n / \tau)} \right] \quad (5)$$

where the numerator is defined as a sum of ‘similarity’ between \mathbf{z}_2^j (i.e. the g_2 transformed output of $\mathbf{t}^{CA,j}$) and a positive set \mathcal{P}_2^j , constructed by identifying the most similar samples using its corresponding *multimodal attention enhanced visual* feature $\mathbf{v}^{CA,j}$. The structure of \mathcal{P}_2^j is similarly defined as:

$$\mathcal{P}_2^j = \{A^{text}(T^j, a), T^k | k \in Nbr_K(\mathbf{z}_1^j \cdot \mathbf{z}_1^i), \forall (\mathbf{s}^i, l^i) \in \mathcal{D} \setminus \{(\mathbf{s}^j, l^j)\}, l^k == l^j, a \in \mathcal{A}^{text}\} \quad (6)$$

where $A^{text}(T^j, a)$ obtains the augmented version of T^j , parameterized by a sampled from a pre-defined set of augmentation transformations in \mathcal{A}^{text} and \mathcal{N}_2^j represents the complement of \mathcal{P}_2^j that does not include samples with the same annotation l^j and their augmentations. Both models $g_1()$ and $g_2()$ are initialized independently by learning in the uni-mode environments and then co-training process proceeds by alternatively optimizing \mathcal{L}_1 (Eqn. 3) and \mathcal{L}_2 (Eqn. 5). In all our experiments, we have used $K = 5$. More about these implementation details will be discussed in Section 4.2.

3.5 Fake News Classification

After the functions g_1 and g_2 are learned, \mathbf{s}^j is represented using its joint multimodal representation $\mathbf{z}^{j, Merged} = [\mathbf{z}_1^j, \mathbf{z}_2^j]$ and is fed into a stack of Fully Connected (FC) layers for classification. In order to address the issue of overfitting,





Image	Text	Ground Truth label	Image-mode Classifier Prediction	Text-mode Classifier Prediction	Multi-mode Classifier Prediction
	"a fish in the new England aquarium"	True	True	True	True
	"other discussions"	False Connection	True	True	False Connection
	"cutest baby cow ive seen in my head all day and just enjoy destiny for what it is"	Misleading Content	Manipulated Content	True	False Connection
	"three corgis larping at the beach"	True	True	Manipulated Content	True

Fig. 3. Example results of the proposed method in 6-way classification task of Fakeddit Dataset [21]

Table 1. Results in Twitter Dataset [4]

Method	Accuracy	Real News F1-Score	Fake News F1-Score
Text-mode Classifier	0.62	0.61	0.64
Visual-mode Classifier	0.64	0.67	0.63
Neural Talk [28]	0.61	0.63	0.59
VQA [2]	0.63	0.61	0.65
EANN [30]	0.65	0.62	0.66
att-RNN [13]	0.66	0.68	0.65
MVAE [15]	0.75	0.76	0.73
Spotfake [26]	0.77	0.70	0.82
Proposed Method	0.82	0.80	0.85

dropout-based regularization is employed, which randomly chooses a percentage κ of hidden units during the updating step. A scaled version of the learned weight ($wt_{sc} = \kappa \cdot wt$) without applying the dropout, is used at the inference step. The standard back propagation algorithm is employed to update FC layer weight parameters. The activation of the last FC layer is fed into a softmax layer to obtain the probabilistic class membership scores.

4 Experiments

In this section, we will discuss the experimental details and the performance of the proposed method using state-of-the-art datasets.

4.1 Dataset Description

While publicly available dataset to evaluate the multimodal fake news detection techniques are relatively rare, in this paper, we use two datasets, which are comprised of multimodal social media contents, which have been popularly used in the research community: (1) Twitter Dataset, as part of MediaEval [4], which was released for evaluating methods for detecting fake multimedia content in social media; (2) the large scale Fakeddit Dataset [21] that was collected using pushshift API to capture samples from up to six different categories of information disorder. Each sample in the Twitter dataset is represented using a

short Twitter message along with the visual and social context information. It has around 17,000 unique tweets discussed on different events and the authors provide the development/test dataset splits with no overlap of events. The development set has 9,000 fake news contents and 6,000 real news contents. The test set has 2,000 tweets. We use the training collection to build the model and the test collection is used for evaluation. The Fakeddit dataset has 1,000,000 samples from up to 6 different categories. Authors provide the ground truth labels for binary fake/real classification as well as more fine-grained categorization of 3 and 6 classes, respectively. While several metadata attributes are also available, which includes up- and down-votes of postings, the number of comments, up- and down-vote score for each comment, to ensure generalization across different data platforms, the proposed method in this work uses only the post content (text and visual) to analyze its veracity. For our experiments we adopt a similar pre-processing technique as in [21] to remove samples which may not have provided information using both modes (text and image) and use the remaining 560,622 samples for training, 58,972 for validation, and 58,954 of the Fakeddit dataset for testing. Each subreddit is labeled with one 2-way, 3-way, and 6-way label. This helps in both high-level and fine-grained fake news classification tasks. The 2-way classification determines whether a sample is fake or true. The 3-way classification determines whether a sample is completely true, the sample is fake and contains text that is true (i.e. direct quotes from propaganda posters), or the sample is fake with false text. The 6-way classification labels are : True, Satire/Parody, Misleading Content, Imposter Content, False Content, and Manipulated Content.

4.2 Implementation Details

For each sample, the text component and their respective images are pre-processed to ensure a uniform size specification. For the text mode, the input length is fixed as 20 tokens, which was decided based on the average length of the text data components in both the datasets. For the image mode, all the images are resized to $224 \times 224 \times 3$. Each image is split into 16×16 blocks (i.e. $B = 16$) and we set $L = 15$. We have chosen $M = 2$ layers in *Attention Aware Image Description* module F_v^{SA} , the *Attention Aware Text Description* module F_t^{SA} , and also in *Multimodal Feature Representation with Attention Aware Fusion* module F^{CA} . The final *Fake News Classifier* is trained with Adam optimizer ($LR = 10^{-5}$) [17] and batch size 128. At the initialization stage of multimodal co-training, both g_1 (and g_2) are initialized independently by identifying the \mathcal{P}_1^j (and \mathcal{P}_2^j) and \mathcal{N}_1^j (and \mathcal{N}_2^j) using the existing \mathbf{z}_1^j (and \mathbf{z}_2^j). Then during alteration, each model (g_1 and g_2) is trained for 50 epochs using the positive sets from the other view. For optimization, we use Adam with 10^{-3} learning rate and 10^{-5} weight decay. For the visual data augmentation to build \mathcal{A}^{vis} , we apply nine crops (center crop plus four corners, with horizontal flipping)) to the visual component of each sample. For the text data augmentation to build \mathcal{A}^{text} , we apply 2 Synonym Replacements, 2 Random Insertions, 2 Random Deletions, and 3 Random Swaps to the text component of each sample.

Table 2. Results in Fakeddit Dataset [21]

Method	2-way		3-way		6-way	
	Validation Accuracy	Test Accuracy	Validation Accuracy	Test Accuracy	Validation Accuracy	Test Accuracy
Text-mode Classifier [21]	0.87	0.86	0.86	0.86	0.76	0.77
Image-mode Classifier [21]	0.74	0.74	0.73	0.73	0.64	0.65
Multimode Classifier [21]	0.87	0.87	0.84	0.84	0.81	0.82
Text-mode Classifier [18]	0.88	0.88	-	-	-	-
Image-mode (Inception V3) Classifier [18]	0.81	0.82	-	-	-	-
Multimode Classifier [18]	0.91	0.91	-	-	-	-
Proposed Method (Text mode)	0.92	0.91	0.87	0.85	0.81	0.80
Proposed Method (Image mode)	0.78	0.76	0.75	0.74	0.69	0.68
Proposed Multimode Classifier	0.94	0.93	0.90	0.89	0.85	0.85

4.3 Results

Table 4 compares the performance of the proposed method against several state-of-the-art algorithms in Twitter dataset using Accuracy, Precision, Recall, and F1 scores (harmonic mean of Precision and Recall). While the proposed multimodal approach demonstrates a significantly improved performance compared to the existing methods in the uni-mode environment, visual feature reports a more reliable performance compared to its text-based counterpart. For the Visual mode classifier, we use the attention aware image representation technique described in section 3.1 and derive a uni-mode implementation of a supervised contrastive loss based learning. More specifically, in the uni-mode environment, the positive set (Eqn 4) is identified by the existing function \mathbf{g}_1 (instead of \mathbf{g}_2 in the multimodal environment) for computing the loss value (Eqn 3). For the text-based classifier also, we adopt a similar approach and leverage the positive set (Eqn 6) identified by the existing function \mathbf{g}_2 for computing the loss value (Eqn 5).

The proposed method is also compared against several state-of-the-art methods including Visual Question Answering (VQA) [2], Neural Talk [28], Event Adversarial Neural Network (EANN) [30] att-RNN [13] and Multimodal Variational Autoencoder (MVAE) [15]. To ensure a fair comparison, we adopt the approach followed by Khatar et al. [15] and build similar architectures for Visual Question Answering (VQA) [2], Neural Talk [28]. The table compares the performance with their two mode frameworks, which do not utilize the social context information. We note that among the other multimodal models, while Spofake and MVAE show improved performance, att-RNN reports better performance compared to EANN specifically in detecting the real-news sub-collection and thereby demonstrates effectiveness of attention mechanism. As observed in the table, by a hierarchical analysis of the local semantics within/across different

modes and a novel co-training mechanism leveraging the supervised contrastive loss in a multimodal environment, the proposed method shows a significantly better performance compared to MVAE by reporting around 6% improvement in accuracy and 9% improvement in F1-score.

The performance of the proposed method using Fakeddit dataset [21] is detailed in Table 2, where the evaluation is performed in multiple steps and the results are reported in both uni-mode and multimode environments. Armin et al. [18] report the performance only using the 2-way annotations of the dataset. As mentioned by the authors of [18], different visual encoders (including VGG-19) were evaluated and Inception-v3 provided the best results for their framework. Therefore, Armin et al. [18] report their best performance using Inception-v3. While leveraging a more discriminative encoder may be a tool for improving the performance for any methods including ours, the objective in this work was to analyze the effectiveness of the proposed multimodal analytical framework, without relying on any specific encoder to boost up the performance. VGG-19 is one of the most popular visual encoders employed by different models developed for this problem scenario, we have used it in our work. By comparing Row 3 with 4, Row 6 with 7, and Row 9 with 10 of Table 2, we observe that the text-mode representation of the news content is most effective in isolation. Also, the uni-mode classifiers (both text and visual) designed in this work, significantly outperform their respective configurations reported by [21] and [18]. Finally the performance of the proposed multimodal classifier that leverages self- and cross-modal attentions at multiple layers of details and learns a discriminative joint representation via multimodal co-training, demonstrates a significant promise in improving the overall identification performance. By observing the accuracy scores reported in Row 11 and comparing them against its corresponding baselines in Row 5 and 8, we find that the proposed method achieves 2 – 5% improvements in accuracy score across various testing environments (2-, 3-, and 6-way). In fact, in the more complex 3- and 6-way problem settings, the proposed method enhances the performance by respectively reporting a more reliable 89% (compared to 84% as reported by [21]) and 85% (compared to 82% as reported by [21]) accuracy scores in the test subcollections. This analysis thus clearly proves beneficial and highlights the intrinsic multimodal nature of the problem setting. Some example results are shown in Figure 3.

5 Conclusion

In this paper we propose a novel multimodal fake-news identification model. To capture a detailed relationship across multiple visual regions and also their correlation with the text component of the input news content, its image component is represented using a set of features describing its potential object regions and its text component as a sequence of words. A pair of mode-specific independent branches of self-attention layers, followed by an attention aware cross-model fusion module learns an initial joint representation by specifically highlighting the correlation between each word and each image region. The proposed multimodal co-training scheme employs an effective formulation of the supervised

contrastive loss based optimization process, which utilizes the complementary category relevant information from different mode-specific data representations to derive an enhanced joint descriptor with improved discriminative capacity. Experiments were performed on two large scale public datasets with news articles with varied characteristics. The consistently improved performances across various experiment settings clearly demonstrate the feasibility of our approach over existing methods. In future we would also like to leverage the access to other metadata (e.g. publishing resource, pattern of responses from viewers, which may be beneficial to evaluate the veracity of content more accurately).

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. *Journal of economic perspectives* **31**(2), 211–36 (2017)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2425–2433 (2015)
3. Bhattacharjee, S.D., Talukder, A., Balantrapu, B.V.: Active learning based news veracity detection with feature weighting and deep-shallow fusion. In: *2017 IEEE International Conference on Big Data (Big Data)*. pp. 556–565. IEEE (2017)
4. Boididou, C., Papadopoulos, S., Dang-Nguyen, D.T., Boato, G., Riegler, M., Middleton, S.E., Petlund, A., Kompatsiaris, Y.: Verifying multimedia use at mediaeval 2016
5. Cui, L., Wang, S., Lee, D.: Same: sentiment-aware multi-modal embedding for detecting fake news. In: *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. pp. 41–48 (2019)
6. Davidson, T., Warmusley, D., Macy, M., Weber, I.: Automated hate speech detection and the problem of offensive language. In: *Proceedings of the International AAAI Conference on Web and Social Media*. vol. 11 (2017)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
8. Ghanem, B., Rosso, P., Rangel, F.: An emotional analysis of false information in social media and news articles. *ACM Transactions on Internet Technology (TOIT)* **20**(2), 1–18 (2020)
9. Giachanou, A., Rissola, E.A., Ghanem, B., Crestani, F., Rosso, P.: The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers. In: *International Conference on Applications of Natural Language to Information Systems*. pp. 181–192. Springer (2020)
10. Giachanou, A., Rosso, P., Crestani, F.: Leveraging emotional signals for credibility detection. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 877–880 (2019)
11. Giachanou, A., Zhang, G., Rosso, P.: Multimodal fake news detection with textual, visual and semantic information. In: *International Conference on Text, Speech, and Dialogue*. pp. 30–38. Springer (2020)
12. Hoffer, E., Ailon, N.: Deep metric learning using triplet network. In: *International workshop on similarity-based pattern recognition*. pp. 84–92. Springer (2015)
13. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: *Proceedings of the 25th ACM international conference on Multimedia*. pp. 795–816 (2017)

14. Kambhatla, N., Leen, T.K.: Dimension reduction by local principal component analysis. *Neural computation* **9**(7), 1493–1516 (1997)
15. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: Mvae: Multimodal variational autoencoder for fake news detection. In: *The World Wide Web Conference*. pp. 2915–2921 (2019)
16. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. *arXiv preprint arXiv:2004.11362* (2020)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
18. Kirchknopf, A., Slijepcevic, D., Zeppelzauer, M.: Multimodal detection of information disorder from social media. *arXiv preprint arXiv:2105.15165* (2021)
19. Lago, F., Phan, Q.T., Boato, G.: Visual and textual analysis for image trustworthiness assessment within online news. *Security and Communication Networks* **2019** (2019)
20. Ma, J., Gao, W., Wong, K.F.: Rumor detection on twitter with tree-structured recursive neural networks. *Association for Computational Linguistics* (2018)
21. Nakamura, K., Levy, S., Wang, W.Y.: r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854* (2019)
22. Popat, K., Mukherjee, S., Strötgen, J., Weikum, G.: Credibility assessment of textual claims on the web. In: *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. pp. 2173–2178 (2016)
23. Popat, K., Mukherjee, S., Yates, A., Weikum, G.: Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018)
24. Shu, K., Wang, S., Liu, H.: Understanding user profiles on social media for fake news detection. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. pp. 430–435. *IEEE* (2018)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
26. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. pp. 39–47. *IEEE* (2019)
27. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. *International Journal of Computer Vision* **104**(2), 154–171 (2013)
28. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 3156–3164 (2015)
29. Wang, W.Y.: ” liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017)
30. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: Eann: Event adversarial neural networks for multi-modal fake news detection. In: *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. pp. 849–857 (2018)
31. Wang, Y., Yang, H., Qian, X., Ma, L., Lu, J., Li, B., Fan, X.: Position focused attention network for image-text matching. *arXiv preprint arXiv:1907.09748* (2019)
32. Zlatkova, D., Nakov, P., Koychev, I.: Fact-checking meets fauxtography: Verifying claims about images. *arXiv preprint arXiv:1908.11722* (2019)