Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang
Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS
University of Chinese Academy of Sciences
Beijing 100049, China
{jinzhiwei,caojuan,guohan,zhyd}@ict.ac.cn

Jiebo Luo
Department of Computer Science
University of Rochester
Rochester, NY 14627, USA
jluo@cs.rochester.edu

ABSTRACT

Microblogs have become popular media for news propagation in recent years. Meanwhile, numerous rumors and fake news also bloom and spread wildly on the open social media platforms. Without verification, they could seriously jeopardize the credibility of microblogs. We observe that an increasing number of users are using images and videos to post news in addition to texts. Tweets or microblogs are commonly composed of text, image and social context. In this paper, we propose a novel Recurrent Neural Network with an attention mechanism (att-RNN) to fuse multimodal features for effective rumor detection. In this end-to-end network, image features are incorporated into the joint features of text and social context, which are obtained with an LSTM (Long-Short Term Memory) network, to produce a reliable fused classification. The neural attention from the outputs of the LSTM is utilized when fusing with the visual features. Extensive experiments are conducted on two multimedia rumor datasets collected from Weibo and Twitter. The results demonstrate the effectiveness of the proposed end-to-end att-RNN in detecting rumors with multimodal contents.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; *Social networks*; • **Computing methodologies** → *Computer vision*;

KEYWORDS

Rumor detection, multimodal fusion, LSTM, attention mechanism, microblog

ACM Reference format:
Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang and Jiebo Luo. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *Proceedings of MM'17, October 23–27, 2017, Mountain View, CA, USA.*, , 9 pages.
DOI: https://doi.org/10.1145/3123266.3123454

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10...$15.00
DOI: https://doi.org/10.1145/3123266.3123454

Figure 1: Some rumor tweets from Twitter.

1 INTRODUCTION

Microblogs, including Twitter and Chinese Weibo, have become important news media and public opinion field in various events recently. For example, during the 2016 U.S. presidential election, candidates and their supporters were actively involved on Twitter to do campaigns and express their opinions. The convenience and openness of microblogs have also fostered various false rumors and fake news, which have become a serious public concern recently. According to the check of Snopes.com, as many as 529 different rumors pertaining Donald Trump and Hillary Clinton were spreading on social media during the election [12], which could have impacts on the election. To increase the credibility of information on mircoblogs and prevent the spreading of fake contents, it is crucial to detect rumors automatically on microblogs.

In addition to texts in tweets, images and videos have gained popularity on microblogs recently. Compared with texts, images can depict visual contents and thus attract much more attention [16]. With rich visual information, they could also be helpful in distinguishing rumors. Figure 1 shows several examples of rumor tweets from Twitter to presents the text and image in each tweet. We can observe that: in the left example, both image and the text indicate it is probably a false rumor; in the middle example, we can tell from the text that it is probably false, although it is hard to tell from the image; in the right example, on the contrary, it is hard to tell the veracity from the text, but the likely-manipulated image suggests it is probably false. We also find microblogs

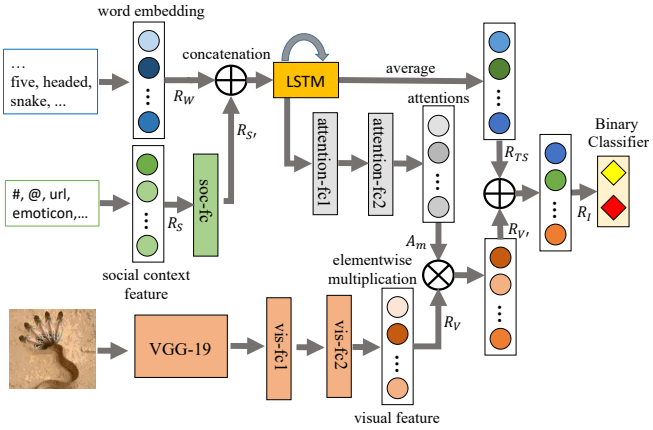

Figure 2: The network structure of proposed multimodal att-RNN. It has three main part: the RNN fusing textual and social context features at the top branch, the visual sub-network at bottom branch, and the neural-level attention from RNN to visual features.

provide us with rich social context, such as hash-tag topics in tweets and retweetings. We aim to fuse different modalities of contents on mircoblogs for detecting rumors in this paper.

Most existing approaches on automatic rumor detection are based on text and social context. Classification-based methods [7, 16, 31] and graph-based optimization methods [10, 13, 14] are deployed to verify online textual posts as false or real based on manually crafted textual and social context features. Only a few recent studies make a first attempt to detect rumors based on multimedia content [9, 15, 16]. Visual features utilized in these works are also crafted manually, which are basically fused with existing features via feature concatenation (early fusion) or averaging results (late fusion). On one hand, hand-crafted features in existing works are limited to learn complicated and scalable textual or visual features. On the other hand, existing fusing methods are quite preliminary which could fail to effectively combine the benefits from different modalities.

Considering these limitations and our motivation to leverage multimodal contents, we propose an end-to-end RNN with attention mechanism to fuse features from text, image and social context for the rumor detection task. Deep neutral networks are proved to be effective in learning accurate textual or visual representations [26, 29]. In the proposed model (Figure 2), we use an RNN to learn the joint representations of text and social context. Image visual features, represented with a pre-trained deep CNN, are then fused with them. We employ the attention mechanism in the model to capture the relations between visual features and joint textual/social features. The contributions of this paper are three folds:

(1) We incorporate multimodal contents on social networks to solve the challenging rumor detection problem. Instead of traditional manually crafted features, textual, visual and social context contents are represented via deep neutral networks.
(2) We propose an innovative RNN with an attention mechanism (att-RNN) for effective multimodal feature fusion. This network fuses features from three modalities and utilizes the attention mechanism for feature alignment.
(3) To validate our model against competing algorithms, we evaluate att-RNN on two multimedia datasets collected from Weibo and Twitter, respectively. The results show that att-RNN achieves the best performance on both datasets, in comparison with exiting feature-based methods and state-of-the-art neural network models.

2 RELATED WORK

In social psychology literature, a rumor is defined as a story or a statement the truth value of which is unverified or deliberately false [1]. Most existing studies solve the automatic rumor detection in feature-based approaches. A wide range of features for detecting rumors are presented in literatures, which could be roughly summarized into three categories: text features, social context features and image features.

Text features represent textual tweets with statistics or semantics. Statistical text features capture prominent statistics in tweets, such as count of word, capitalized characters and punctuation [7]. Semantic text features represent abstract semantics of texts, which include sentiment scores [7] and opinion words [19]. Some works [10, 13] use bag-of-words text features to reveal inter-tweet relations. Features based on topic model, such as LDA are also employed to represent high-level abstract semantics in the rumor detection task [14, 31]. To overcome the limitation of these manually crafted features, Ma *et al.* [21] explore the possibility to represent tweets in an event with deep neural networks. They use RNN to learn the representation of tweets in a time series.

The social connection feature of microblogs give birth to rich social context for posts. They are widely used in rumor detection on social media [31]. Some social context features are designed to capture the interactions on mircoblogs, including the number of retweeting and replying. Other social context features are derived from the feature of social media, such as the usage of hash-tag topics(#), mentions(@) and URLs.

Only a few recent studies aim to verify the credibility of multimedia content in addition to pure text. Morris *et al.* [24] release a survey result that user profile image could indicate the credibility of the user's posts. For images attached in tweets, some basic features are proposed in literature [10, 31]. Aiming to automatically predict whether a tweet that shares multimedia content is fake or real, Boididou *et al.* [5] proposed the Verifying Multimedia Use task which took place as part of the MediaEval benchmark in 2015 and 2016. Text and image forensics features are extracted as baseline features for this task [15]. In the latest work, Jin *et al.* [16] propose several image features based on the visual appearance and statistics

of images in tweets. Combined with textual features, these novel image features prove to be quite effective. However, these features are still hand-crafted and the fusion of text and image features are very simple early feature concatenation and late result fusion.

Deep neural networks have proved to be able to learn accurate image and sentence representations than traditional hand-crafted features. Specifically, convolutional neural networks (CNN) have shown their powerful abilities on image representation [18, 27] while recurrent neural networks (RNN) are widely employed recently in sentence representation [3, 22]. Inspired by their successes, recent multimodal representation learning use neural networks to fuse multimodal features in many applications, such as visual question answering [2] and image captioning [17, 30, 32]. It shows the capability of deep neural networks in bridging the "semantic gap" in multimodal data analytics. Due to the discrepancy in task settings, existing multimodal fusion models have different input feature sets and optimization assumptions, compared with rumor detection task.

Considering the available multimodal features from microblogs and the neural network fusion methods, we propose to fuse textual, social context and visual content with RNN for the rumor detection task. Our proposed model leverages the feature learning and fusing power of deep neural networks and is expected to improve the rumor detection performance.

3 MODEL

Tweets on microblogs provide information in different modalities: tweet text, attached image and surrounding social context. In this paper we aim to comprehensively utilize these multimodal information to determine whether a tweet is a false rumor or not. Based on this basic idea, we propose a novel deep neural network with attention mechanism (att-RNN) to capture intrinsic relations among textual, social context and visual features of a tweet instance. The detailed descriptions of the proposed model are presented in this section.

3.1 Model Overview

We define a tweet instance $I = \{T, S, V\}$ as a tuple representing three different modalities of contents: the textual content T, the social context S, and the visual content V. The proposed model takes features from these modalities (R_T, R_S and R_V), and aims to learn a reliable representation R_I as the aggregation of T, S and V for the given tweet I. Firstly, we fuse the text and social context with a RNN, which generates a joint representation R_{TS} for these two modalities. For a visual feature R_V, it is gained through a deep CNN. We use the attention from the RNN's output at each time step to further refine R_V. In the last step, both R_{TS} and attention-aggregated R'_V are concatenated as the final multimodal representation R_I, upon which a binary classifier is followed to distinguish the tweet instance as real or false.

The overall structure of the proposed att-RNN is illustrated in Figure 2. It has three main parts: 1) an RNN sub-network (the top branch in Figure 2) which learns the joint representation of textual and social context features ; 2) a visual sub-network (the bottom branch in Figure 2) which generates a visual representation; and 3) the neural-level attention part which uses the output of RNN to align visual features. This network is an end-to-end model for detecting rumors with multimodal contents as input.

3.2 LSTM Networks

We employ an RNN with Long Short-Term Memory (LSTM) units to learn the joint representation of text and social context in the proposed model. An RNN is a type of feed-forward neural network that can be used to model variable-length sequential information such as sentences or time series. Given an input sequence $(x_1, x_2, ..., x_M)$, a basic RNN model updates the hidden states $(h_1, h_2, ..., h_M)$ and generates the output vector $(y_1, y_2, ..., y_M)$. Here M depends on the length of input. The current hidden state is estimated using a recurrent unit. In particular, the recurrent unit takes the last hidden state and the current input to produce the current hidden state.

To deal with vanishing or exploding gradients [4, 25] in learning long-distance temporal dependencies, LSTM extends the basic RNN by storing information over long time periods in elaborately designed memory units. To be specific, the reading and writing memory cell c in LSTM is controlled by a group of sigmoid gates: an input gate i, an output gate o and a forget gate f. For each time step m, the LSTM cell receives inputs from the current input x_m, the previous hidden state h_m and the previous memory cell c_m. These gates are updated as follows [8, 11].

$$i_m = \sigma(W_{xi}x_m + W_{hi}h_{m-1} + b_i) \tag{1}$$

$$f_m = \sigma(W_{xf}x_m + W_{hf}h_{m-1} + b_f) \tag{2}$$

$$o_m = \sigma(W_{xo}x_m + W_{ho}h_{m-1} + b_o) \tag{3}$$

$$g_m = \phi(W_{xc}x_m + W_{hc}h_{m-1} + b_c) \tag{4}$$

$$c_m = f_m \odot c_{m-1} + i_m \odot g_m \tag{5}$$

$$h_m = o_m \odot \phi(c_t) \tag{6}$$

where W_* are weight matrices for corresponding gates and b_* are bias terms, which are learned from the network. σ is the sigmoid activation function: $\sigma(x) = 1/(1 + exp(-x))$. ϕ is the hyperbolic tangent function: $\phi(x) = (exp(x) - exp(-x)/(exp(x) + exp(-x))$. $\odot$ denotes the element-wise multiplication between two vectors. The input gate i decides the degree to which the new memory is added to the memory cell. The forget gate f determines the degree to which the existing memory is forgotten. The memory cell c is updated

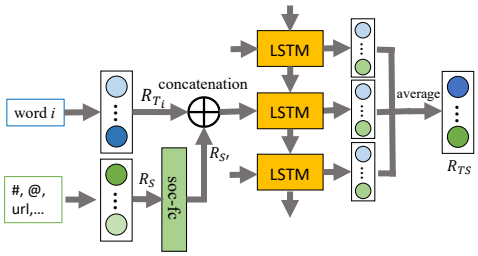

Figure 3: Fusing social context and textual features with LSTM.

by forgetting part of the existing memory and adding new memory g.

3.3 Joint Representation of Text and Social Context

A text content is a sequential list of words it contains: $T = \{T_1, T_2, ..., T_n\}$ (n is the number of words in a text). Each word $T_i \in T$ in the text is represented as a word embedding vector. The embedding vector for each word is obtained with a deep network which is unsupervisedly pre-trained on given dataset.

Traditional multimodal approaches are normally based on textual and visual/audio features. For the rumor detection task on mircoblogs, however, the social context features are widely used in literatures [7, 16, 31] for effective rumor detection. We assume that incorporating social context into the rumor detection model would have some benefits. The context generated on microblogs, such as hash-tag topic, mention and retweets, as well as some text semantic features, such as the emotional polarity, are utilized to form the initial social context representation $R_S = [s_1, s_2, .., s_k]^T$. k is the dimension of social context features and s_i is the scalar value of the i-th dimension. The social context feature R_S is transformed into a representation $R_{S'}$ which has the same dimension of word embedding vector via a fully-connected layer ("soc-fc" in Figure 3) as follows.

$$R_{S'} = W_{sf} R_S \tag{7}$$

where W_{sf} are weights in the fully-connected layer for dimension transformation.

At each time step, the LSTM takes $R_{T_i S} = [R_{T_i}; R_{S'}]$ as input, which is the concatenation of i-th word embedding R_{T_i} and the transformed social context feature $R_{S'}$. The output neurons of LSTM for each word are averaged to form the joint representation of text and social context R_{TS}. The whole process is illustrated in Figure 3.

3.4 Visual Representation of Image

The visual sub-network (the bottom branch in Figure 2) takes tweet images as input and generates visual neurons as features for images. Its front layers have the same structure as VGG-19 network [27]. We add two 512-neuron fully-connected layers ("vis-fc1" and "vis-fc2" in Figure 2) on top of the second to the last layer of VGG-19 net to generate a 512-neuron visual representation $R_V = [v_1, v_2, ..., v_{512}]^T$ for each image. The whole att-RNN network is trained jointly so that the learned visual neurons determines certain patterns for rumor detection. The visual sub-network could be first fine-tuned with auxiliary dataset. During the joint training with the LSTM sub-network, however, only parameters of the last two fully-connected layers are updated for more efficient training.

$$R_V = W_{vf_2} \psi(W_{vf_1} R_{V_p}) \tag{8}$$

where R_{V_p} is the visual feature obtained from the pre-trained VGG net, W_{vf_1} are weights in the first fully-connected layer with ReLU activation and W_{vf_2} are weights in the second fully-connected layers with softmax function, and ψ denotes the ReLU activation function.

In this process, visual neurons are not constrained to learn particular concepts in the network. Hopefully, necessary semantic concepts for rumor detection can be captured by them during the jointly training of the whole network given the training data labeled as rumors or non-rumors.

One big challenge of directly utilizing the visual and joint social-textual representation in the model is that one representation will probably overwhelm the other, which results in the biased performance towards this modality. To maximize the benefits of multimodal features, we should jointly learn the alignments in different modalities. In the following part, we introduce an attention mechanism to adjust the visual representations according to the RNN output at each time step and produce the aggregated visual neurons simultaneously.

3.5 Attention for Visual Representation

We assume images would have certain correlations with text/social context in rumor tweets. In order to characterize these relations, we propose a neuron-level attention mechanism for visual features from the guidance of neuron jointly representing text and social context. Attention mechanism has been successfully applied to match textual and visual semantic concepts in recent language-vision tasks [20, 32]. Under the setting of rumor detection, we assume that words of the text content are likely associated with some semantic concepts in the image. Our goal is to automatically find such kind of connections. In particular, the visual neurons having similar semantic meanings with the word should be assigned with more weights.

Our proposed visual-neuron attention mechanism weights the contributions of different neurons for different words. To achieve this goal, we take the output hidden state h_m of the LSTM at each time step as guidance. h_m is connected into a fully-connected layer with non-linearity ReLU function and a fully-connected layer with softmax function to obtain the attention vector $A_m \in \mathbb{R}^{512}$, which has the same dimension as the visual neurons R_V.

$$A_m = W_{af_2} \psi(W_{af_1} h_m) \tag{9}$$

where h_m is hidden state of LSTM at the m-th time step, W_{af_1} and W_{af_2} are weights in the two fully-connectedly

layers, and ψ is the ReLU activation. The correlation between the m-th word in the text and the image can be calculated as follows:

$$a_m = \sum_{i=1}^{512} A_m(i) v_i \quad (10)$$

where $A_m(i)$ is the attention value for the i-th visual neuron.

Following this attention mechanism, the attention vector A_m generated by LSTM in the joint feature learning of text and social context can decide which visual neurons should be given more emphasis. The final visual representation is the a set of affinity values for each words: $R_{V'} = [a_1, a_2, ..., a_n]$ (n is the number of words in the given text).

It should be mentioned that high-level visual semantics in rumor detection task can be very hard to be identified, compared with object level semantics in traditional visual recognition tasks. There is no mechanism that explicitly guarantees the learning of this matching relation in the attention model. But we still assume that training with such attention mechanism could discover some relations implicitly and improve the feature alignment. Its effectiveness in practice will be validated in the experiment section.

3.6 Model Learning

Till now, we have obtained a joint representation R_{TS} for text and social context and an attention-aggregated visual representation $R_{V'}$. These two features are concatenated to form the multimodal representation $R_I = [R_{TS}; R_{V'}]$ for the given tweet, which is fed into a softmax layer before computing its loss for the rumor detection goal. We employ the cross-entropy to define the loss of m-th tweet as follows.

$$p(R_I^m) = softmax(W_s R_I^m) \quad (11)$$

$$L(R_I^m) = -[l^m \log p(R_I^m) + (1 - l^m) \log p(1 - R_I^m)] \quad (12)$$

where R_I^m is the multimodal representation of the m-th tweet instance, W_s is the parameters in the softmax layer for a linear model, and l^m denotes the ground truth label for the m-th instance with 1 representing false rumor tweets and 0 representing real ones.

In summary, our proposed multimodal att-RNN takes inputting training data $I = \{T, S, V\}$ with contents from three different modalities: text, social context and image. It outputs the prediction label for each instance to indicate it as rumor or non-rumor. The whole model is trained end-to-end with batched Stochastic Gradient Descent to minimize the loss function:

$$L = -\frac{1}{N} \sum_{m=1}^{N} [l^m \log p(R_I^m) + (1 - l^m) \log p(1 - R_I^m)] \quad (13)$$

where N is the total number of instances. The parameters in RNN sub-network, visual sub-network, and attention layers will be learned simultaneously during the optimization process.

Table 1: Statistics of two datasets

Statistics		Weibo	Twitter
Training Set	Rumor	3749	7334
	Non-rumor	3783	5599
Testing Set	Rumor	1000	564
	Non-rumor	996	427
All		9528	13924

4 EXPERIMENTS

To the best of our knowledge, this work is the first study on fusing multimodal features with neural networks for the rumor detection task. We build a number of competing baselines to evaluate the performance. In addition to traditional feature-based single-modal methods and simple fusion methods for rumor detection, we also investigate several state-of-the-art language-vision models. Extensive experiments are conducted based on two real-world datasets collected from Weibo and Twitter to demonstrate the effectiveness of our proposed att-RNN for the rumor detection problem.

4.1 Datasets

There are very a few standard multimedia rumor detection datasets available. In addition to the dataset presented in the MediaEval Verifying Multimedia Use benchmark [5], we build a new dataset for the task.

4.1.1 Weibo Dataset. In order to provide a fair evaluation on our proposed method, we collect a dataset from Weibo which have objective ground-truth labels. Specifically, we crawl all the verified false rumor posts from May, 2012 to January, 2016 on the official rumor debunking system of Weibo. This system encourages common users to report suspicious tweets on Weibo. A committee composed of reputable users then would examine the cases and verify them as false or real. This system actually serves as an authoritative source to collect rumor tweets in literatures [21, 31]. For the non-rumor tweets, we use the tweets verified by Xinhua News Agency, an authoritative news agency in China.

Unlike most existing datasets that are focused on only text content, we aim to build a multimedia dataset with images included. We collect the original tweet texts, attached images and available surrounding social contexts from the rumor and non-rumor sources. The raw set contains about 40k tweets with images, after removing text-only tweets. Tweets on wild social media are usually redundant and noisy. We remove duplicated images from the raw set with a near-duplicated image detection algorithm based on locality sensitive hashing (LSH) [28]. We also remove very small or long images to maintain a good quality. For the training/testing data split, we carefully split the tweets concerning the same events (events are discovered with a single-pass clustering method [13]), to make sure they are not contained in both training and testing sets at the same time. Otherwise, they may produce misleading results. The training and testing sets contain a number of tweets approximately with a ratio of 8:2.

4.1.2 Twitter Dataset. The Verifying Multimedia Use task at MediaEval [6] aims to automatically detect false multimedia content on social media. The proposed dataset has two part: one is the development set containing about 9,000 rumor and 6,000 non-rumor tweets from 17 rumor-related events; the other is the test set containing about the 2,000 tweets from another batch of 35 rumor-related events. Thus, the tweets in two sets have different coverings of events. For each tweet, the text content, attached image/video and several social context are available in this dataset. Given our focus on image content in this paper, we remove tweets with video attached in the set. We use the development set as the training set and the test set as testing set to keep the same data split scheme as the benchmark. The detailed statistics of two datasets are listed in Table 1.

4.2 Experimental Settings

For the textual feature, we employ the distributed representation for words [23]. For each of the two datasets, we pre-train the Word2Vec model with the whole dataset in an unsupervised fashion with default parameter settings after standard text pre-procession. We obtain a 32-dimensional word embedding feature for each word in the datasets. One reason to choose the word embedding representation instead of one-hot word representation is that insufficient text would lead to poor word features when the vocabulary size is too large in the one-hot representation approach.

For the social context feature, we take the most social features in literatures. Given available data, we extract 16 social features for the Weibo dataset and 18 features for the Twitter dataset, respectively. The shared and different social features are listed in Table 2. Some text semantic features, such as emotional polarity and number of first order pronouns, are also added into the list for their importance in the rumor detection task.

For the visual feature, we use the output of the second to the last layer of a 19-layer VGGNet pre-trained on Imagenet set [27]. The feature size is 4096. The proposed visual subnetwork could be fine-tuned with additional data, but we leave it to the future research as there is no available image set directly related to this task at the moment.

For the joint learning of text and social context with RNN, we implement it with an LSTM. We set the dimension of the hidden layers as 32 and use hyperbolic tangent function as non-linear activation function. We use a batch size of 128 instances in the training of the whole network. In the following experiments, each neural network model is trained for 100 epochs with an early stopping to report the results.

4.3 Performance Comparison

To validate the proposed multimodal fusion model on rumor detection task, we compared it with two groups of baseline methods. The first group is feature-based rumor detection methods, which are widely employed in recent rumor detection literatures:

Table 2: Features for Text-based Method

Shared	Number of exclamation/question mark
	Number of words/characters
	Number of positive/negative words
	Number of first/second/third order of pronoun
	Number of URL/@/#
Weibo	Number of People/Location/Organization
	Sentiment score
Twitter	Contains of happyemo/sademo
	Number of uppercase characters
	Number of retweets

Single textual model. We use a 400-dimension paragraph embedding feature for each tweet. The feature vectors are fed to a logistic classifier to train a rumor detection model.

Single visual model. The 4096-dimension visual features from a pre-trained VGG-Net are used to train a logistic regression model.

Single social context model. We use social features in Table 2 to train logistic classifiers on two datasets.

Early fusion model. Features from three modalities are concatenated and fed into a logistic regression model.

Late fusion model. The average of the prediction scores of single textual, social context and visual models is used as the prediction score of the late fusion model.

Although this work may be the first to employ deep neural networks to fuse multimodal contents for the rumor detection task, it is necessary to show the full advantage of our model by comparing it with the state-of-the-art multimodel fusion methods proposed in other language-vision applications. To this end, our model is compared with a group of multimodal approaches based on neural networks, which are proposed in the related areas, including visual question answering and image captioning. Considering the discrepancy in the task settings, we take the main network structures of these models and make necessary adaption for our task.

VQA Visual question answering aims to answer questions about given images. We adopt the Visual QA model in [2] for our binary classification task. The element-wise multiplication between text and image are replaced with feature concatenation and the multi-class classifier are replaced with a binary classifier. We also change the LSTM to one layer for a fair comparison. The modified algorithm is denoted as VQA*.

NeuralTalk In the application of image captioning, Vinyals *et al.* [30] propose to generate natural sentences describing an image using deep recurrent framework. We follow their main network structure and take the average of the output of RNN at each time step as the joint representation of image and text in tweets. Feeding the representation into two fully-connected layer followed by an entropy loss layer, we adopt the NeuralTalk model (denoted as NeuralTalk*) for the rumor detection problem.

Noticing that none of social context features is utilized in these two algorithms, we also remove the social features

Table 3: The results of different methods on two datasets

Dataset	Method	Accuracy	Rumor			Non-rumor		
			Precision	Recall	F_1	Precision	Recall	F_1
Weibo	Textual	0.592	0.605	0.531	0.566	0.581	0.653	0.615
	Visual	0.608	0.61	0.605	0.607	0.607	0.611	0.609
	Social Context	0.65	0.672	0.591	0.629	0.634	0.71	0.67
	Early Fusion	0.603	0.612	0.567	0.589	0.595	0.639	0.616
	Late Fusion	0.669	0.693	0.611	0.649	0.651	0.728	0.687
	VQA* [2]	0.736	0.797	0.634	0.706	0.695	0.838	0.76
	NeuralTalk* [30]	0.726	0.794	0.613	0.692	0.684	0.84	0.754
	att-RNN w/o social	0.772	0.854	0.656	0.742	0.72	0.889	0.795
	att-RNN	**0.788**	**0.862**	**0.686**	**0.764**	**0.738**	**0.89**	**0.807**
Twitter	Textual	0.532	0.598	0.541	0.568	0.462	0.52	0.489
	Visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	Social Context	0.509	0.566	0.589	0.577	0.426	0.403	0.414
	Early Fusion	0.619	0.727	0.528	0.612	0.542	0.738	0.625
	Late Fusion	0.594	0.661	0.589	0.623	0.526	0.602	0.561
	VQA* [2]	0.631	0.765	0.509	0.611	0.55	**0.794**	0.65
	NeuralTalk* [30]	0.61	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN w/o social	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	att-RNN	**0.682**	**0.78**	**0.615**	**0.689**	**0.603**	0.77	**0.676**

in our att-RNN model for a fair comparison (denoted as "att-RNN w/o social").

Table 3 shows the performance results for all the above models. We can observe that on both datasets, our proposed att-RNN model significantly outperforms all the baseline models. The att-RNN model achieve an overall accuracy of 78.8% on Weibo set and 68.2% on Twitter set, which indicates it can learn the joint features from multiple modalities effectively.

For the feature-based fusion model (early/late fusion), the results are mostly dominated by the modality which has the most prominent detection power. The indicates that simple fusion models probably fail to coordinate features from different feature spaces. Without fusing any social features, two multimodal fusing models based on neural networks (VQA* and NeuralTalk*) still perform better than feature-based fusing methods. This shows they can effectively fusing textual and visual features into a joint representation via the deep networks. By introducing the attention mechanism for feature alignment, the "att-RNN w/o social" can outperform these neural networks models. With additional social features in the network structure, the proposed att-RNN can further improve its performance and reach a reliable multimodal representation.

Specifically, on the Weibo dataset att-RNN boosts the rumor detection accuracy of single modality method from 65% to 78.8% and outperforms the feature fusion method by over 12%. On the Twitter dataset, the accuracy of att-RNN increases from 59.6% to 68.2%, compared with the best single modal method (visual feature). It also outperforms the feature fusion model by over 6%. For the two fusing models based on neural network, VQA* and NeuralTalk*, they are observed to have better performance than feature-based methods on both datasets. These observations prove the advantage of fusing multimodal features via neural networks. Moreover, att-RNN, with or without fusing social features, improves the accuracy of two state-of-the-art multimodal fusing networks by 5% on both datasets. This validates the effectiveness of att-RNN in detecting rumors on microblogs.

4.4 Component Analysis

To further investigate the impact of each modality and the attention mechanism in the proposed model, we design several baselines for comparison, which are simplified variations of att-RNN by removing certain components:

att-RNN w/o attention Proposed att-RNN without the attention mechanism.

att-RNN w/o social Social context is removed from att-RNN, the attention vector is now merely based on text LSTM.

att-RNN w/o social+attention We remove social context and the attention mechanism in proposed model. The remaining structure is just the same as VQA* in the last part.

att-RNN w/o image We remove the visual sub-network and associated attention mechanism. The joint feature of text and social context learned from LSTM is fed into a binary classifier.

att-RNN w/o image+social We train an LSTM with text as input and feed the averaged outputs into a binary classifier. This model can be served as a neural network approach based on only textual content [21].

We report the accuracy and F_1 scores of these baselines on both datasets in Table 4. We can observe that all the components: social context features, visual features and neural-level

Table 4: The results of component analysis on two datasets

Method	Weibo			Twitter		
	Accuracy	Rumor F_1	Non-rumor F_1	Accuracy	Rumor F_1	Non-rumor F_1
att-RNN	**0.788**	**0.764**	**0.807**	**0.682**	**0.689**	**0.676**
w/o attention	0.745	0.71	0.773	0.668	0.68	0.655
w/o social	0.772	0.742	0.795	0.664	0.676	0.651
w/o social+attention	0.736	0.706	0.76	0.631	0.611	0.65
w/o image	0.743	0.708	0.771	0.625	0.642	0.605
w/o image+social	0.721	0.683	0.752	0.613	0.693	0.474

(a) Top rumor examples

(b) Top non-rumor examples

Figure 4: Some top rumors and non-rumors detected by att-RNN from the Weibo set.

Figure 5: Some rumors detected by att-RNN but missed by text-only RNN on the Weibo set.

attention, are all important for achieving the best rumor detection performance by our att-RNN. If we remove one or several of them, the performance would drop by a certain degree. In particular, visual features play a vital role: the accuracy drops by 7% without them on both datasets. Social context features and attention are also very important. On both datasets, the accuracy are 5% lower than att-RNN, without social context and attention.

4.5 Case Study

We provide a qualitative analysis of att-RNN on some successful examples. We rank the detected rumors and non-rumors based on the prediction scores of att-RNN and illustrate three of each class from the Weibo dataset in Figure 4. For simplicity and clarity, only text and image content are presented in these examples. Compared with detected non-rumors, the three rumor examples reveal that att-RNN takes advantage of both textual and visual contents for rumor detection. Compared with non-rumor ones, rumor examples seem to have some rumor patterns in text or image content.

To illustrate the importance of fusing various modalities, we run a cross examination of the two result lists: one is from the att-RNN and the other is from att-RNN with only text as input ("att-RNN w/o image+social"). In particular, we examine the rumors successfully detected by att-RNN but missed by text-only RNN. Figure 5 shows three top-confident examples from the cross examination. Their textual contents show little evidence of rumor patterns, however, the visual contents seem like forged pictures.

5 CONCLUSIONS

We propose an RNN with the attention mechanism to fuse features from text, image and social context for detecting rumors on microblogs. For a given tweet, its text and social context are first fused with an LSTM. The joint representation are then fused with visual features extracted from pre-trained deep CNN. The output of the LSTM at each time step is employed as the neuron-level attention to coordinate visual features during the fusion. Extensive experiments conducted on Weibo and Twitter datasets show that the proposed att-RNN model can effectively detect rumors based on multimedia content, in comparison with existing feature-based methods and several multimodal fusion methods based on neural networks.

ACKNOWLEDGMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0800403, the National Natural Science Foundation of China (61571424, 61525206) and the Beijing Advanced Innovation Center for Imaging Technology under Grant BAICIT- 2016009. Jiebo Luo would like to thank the support from the New York State through the Goergen Institute for Data Science. Zhiwei Jin gratefully thanks the sponsorship from the China Scholarship Council.

REFERENCES

[1] Gordon W Allport and Leo Postman. 1947. The psychology of rumor. (1947).

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* 5, 2 (1994), 157–166.

[5] Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, and Yiannis Kompatsiaris. 2015. Verifying Multimedia Use at MediaEval 2015. In *MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany*.

[6] Christina Boididou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2014. Challenges of computational verification in social multimedia. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*. 743–748.

[7] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web (WWW)*. ACM, 675–684.

[8] Felix A Gers, Nicol N Schraudolph, and Jürgen Schmidhuber. 2002. Learning precise timing with LSTM recurrent networks. *Journal of machine learning research* 3, Aug (2002), 115–143.

[9] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: characterizing and identifying fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 729–736.

[10] Manish Gupta, Peixiang Zhao, and Jiawei Han. 2012. Evaluating Event Credibility on Twitter. In *Proceedings of the SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 153.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[12] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter. In *Social, Cultural, and Behavioral Modeling - 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings*. 14–24.

[13] Zhiwei Jin, Juan Cao, Yu-Gang Jiang, and Yongdong Zhang. 2014. News credibility evaluation on microblog with a hierarchical propagation model. In *2014 IEEE International Conference on Data Mining (ICDM)*. IEEE, 230–239.

[14] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*.

[15] Zhiwei Jin, Juan Cao, Yazi Zhang, and Zhang Yongdong. 2015. MCG-ICT at MediaEval 2015: Verifying Multimedia Use with a Two-Level Classification Model. In *Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*.

[16] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Trans. Multimedia* 19, 3 (2017), 598–608.

[17] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[19] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 1103–1108.

[20] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. *arXiv preprint arXiv:1702.05729* (2017).

[21] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of IJCAI*.

[22] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model.. In *Interspeech*, Vol. 2. 3.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[24] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing?: understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 441–450.

[25] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. *ICML (3)* 28 (2013), 1310–1318.

[26] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229* (2013).

[27] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[28] M. Slaney and M. Casey. 2008. Locality-Sensitive Hashing for Finding Nearest Neighbors. *IEEE Signal Processing Magazine* 25, 2 (March 2008), 128–131.

[29] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[31] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *IEEE International Conference on Data Engineering, ICDE*.

[32] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1008–1017.