2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)

SpotFake: A Multi-modal Framework for Fake News Detection

Shivangi Singhal
IIIT-Delhi
Delhi,India
shivangis@iiitd.ac.in

Rajiv Ratn Shah
IIIT-Delhi
Delhi, India
rajivratn@iiitd.ac.in

Tanmoy Chakraborty
IIIT-Delhi
Delhi, India
tanmoy@iiitd.ac.in

Ponnurangam Kumaraguru
IIIT-Delhi
Delhi, India
pk@iiitd.ac.in

Shin'ichi Satoh
NII
Tokyo, Japan
satoh@nii.ac.jp

BigMM'19

220126 Chia-Chun Ho

Outline

Introduction

Related Works

Methodology

Experiments

Conclusions

SpotFake+

Comments

Introduction

Fake news examples

- A study define fake news "to be news articles that are intentionally and verifiably false, and could mislead readers."
- Moreover, such content is written with the intention to deceive someone.
- The image shown in the news is photo-shopped to make it look similar to the news that is generally features on a popular news channel like CNN.
- This image made people believed that the news is real, but it was later quashed by the victim himself.

Figure 1: An example of fake news that claims that the actor Sylvester Stallone died due to prostate cancer.

Figure 2: The reply from the actor after the spread of the news of his death.

Introduction

Multi-modal features are more beneficial

- The authors did a survey on a sample population that consisted of people in the age group in the age group of 15-50 years.
- 81.4% people can correctly identifying fake news when given multi modalities.
- The survey confirms the fact that multi-modal features are more beneficial in detecting fake news as compared to uni-modal features.

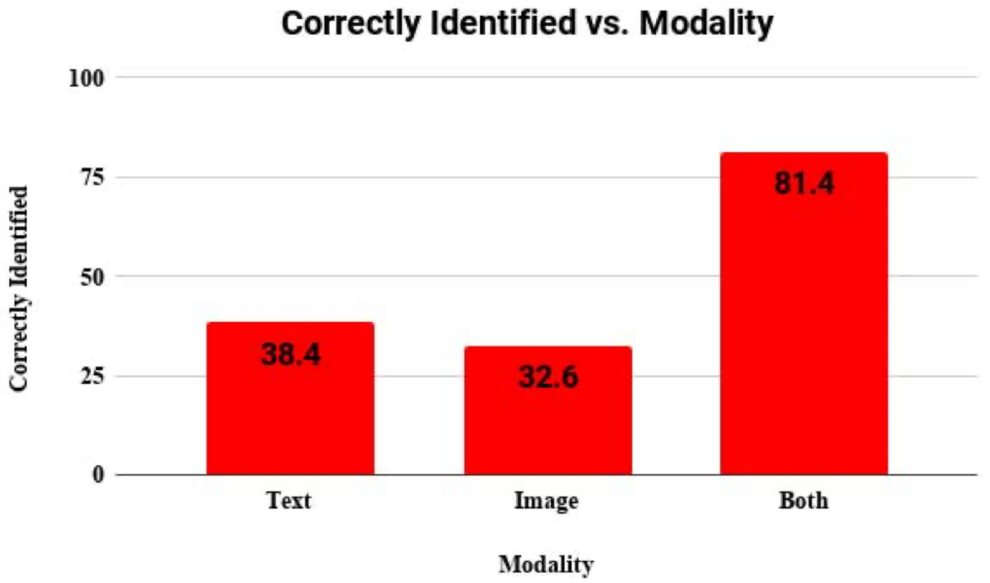
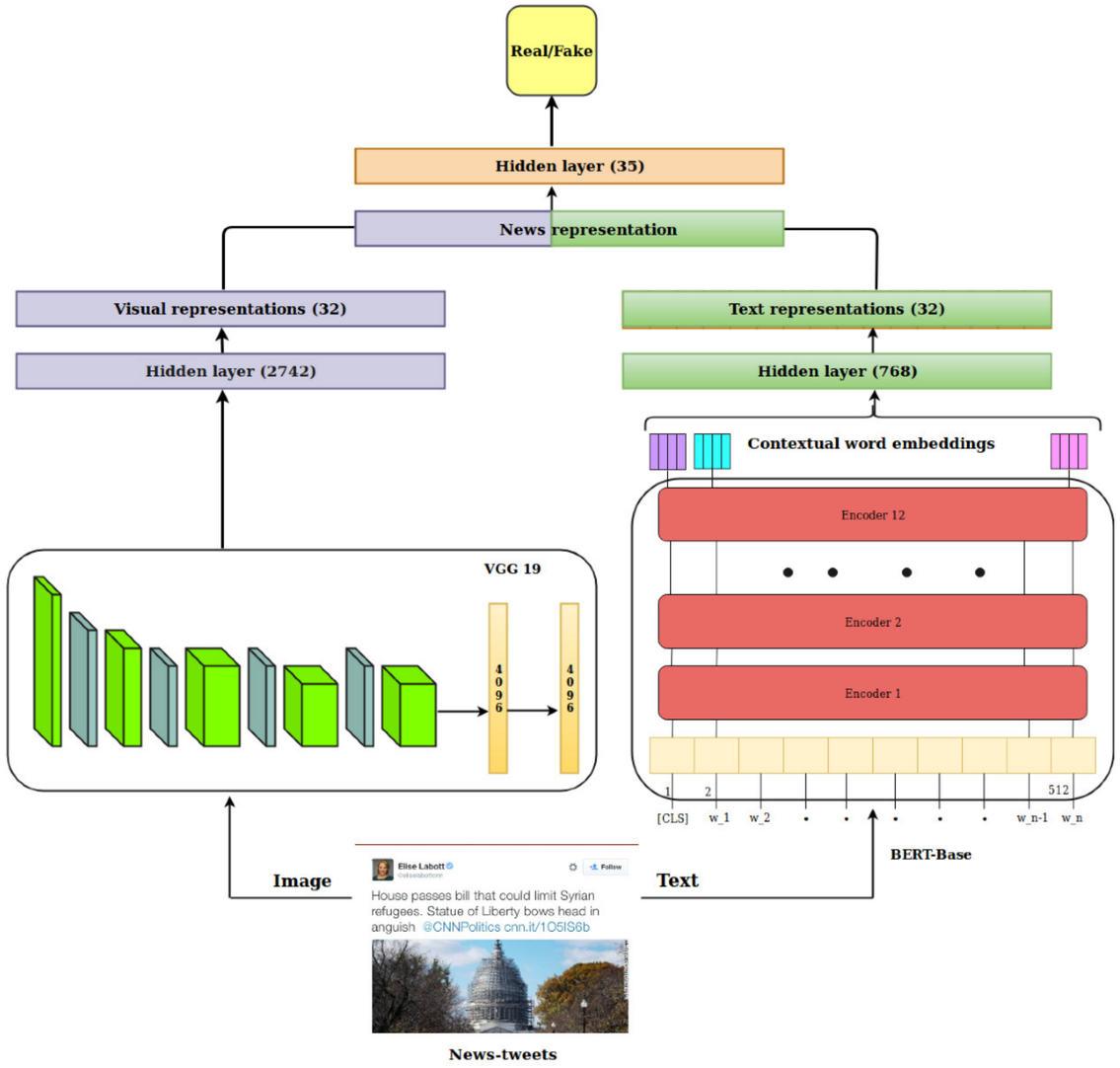


Figure 8: Percentage of people that were successful in identifying fake news when given different modalities.

Introduction

SpotFake

- Proposed SpotFake - a multi-modal framework for fake news detection.
- SpotFake takes into account the two modalities present in an article
  - text and image.



Introduction

Motivations

- Different modalities exhibit different aspects of a news.
- Information derived from different modalities complement each other in detecting the authenticity of news
- Different sources manipulate different modalities based on their expertise.
  - Some people have experience in creating fake news by manipulating images and others may have experience in manipulating modalities such as text, audio and videos.
- Since real-world texts, photos, and videos are complex, contextual information is also important in addition to content information.

Introduction

Contributions

- Proposed architecture aims to detect whether a given news article is real or fake.
- It doesn't take into account any other sub-task in the detection process.
- The prime novelty of SpotFake is
  - incorporate the power of language model BERT to incorporate contextual information.
  - image features are learned from VGG-19 pre-trained on ImageNet dataset.
- Representations from both the modalities are then concatenated together to produce the desired news vector which finally used for classification.

Related Works

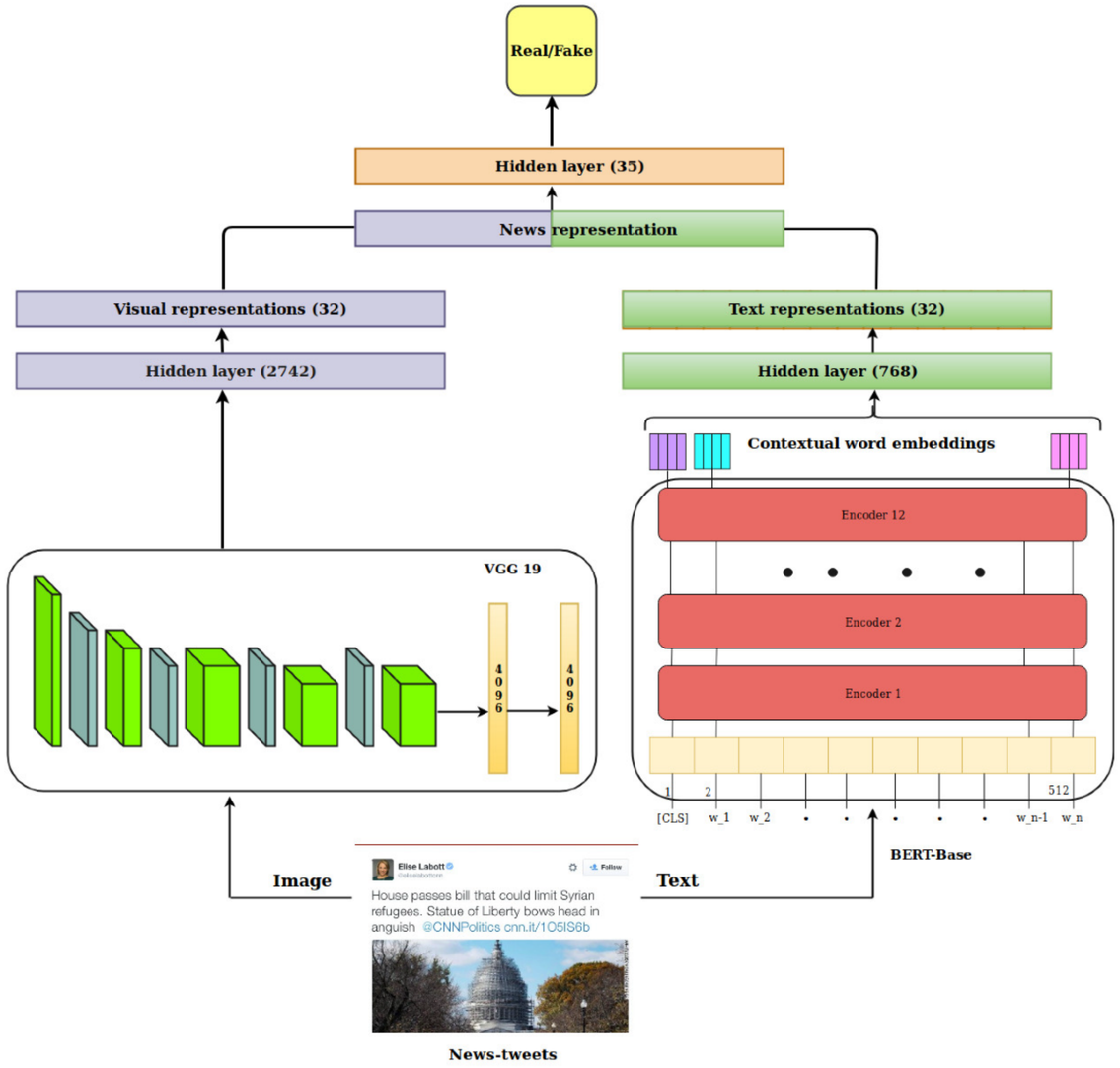
EANN, MVAE

- These multimodal system perform well in detecting fake news, the classifiers have always been trained in tandem with another classifier.
- This increases training and model size overhead, increases training complexity and at times can also hinder the generalizability of the systems due to lack of data for the secondary task.
- To solve such issues, SpotFake takes into consideration features from two different modalities and classifies the sample into real or fake without taking into account any other sub-task.

Methodology

Framework overview

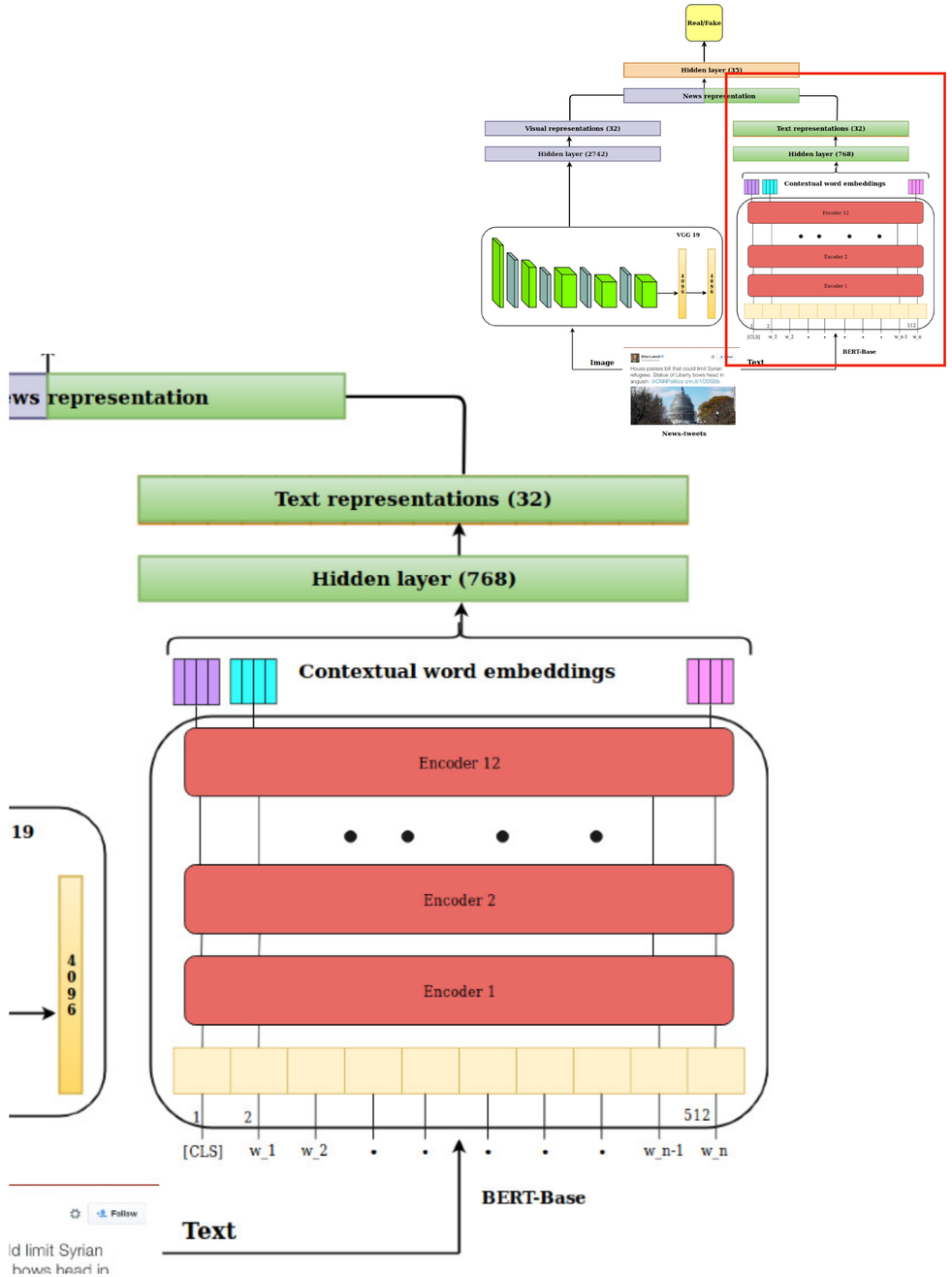
- SpotFake divide into three sub-modules:
  - Textual feature extractor
  - Visual feature extractor
  - Multimodal fusion module



Methodology

Textual Feature Extractor

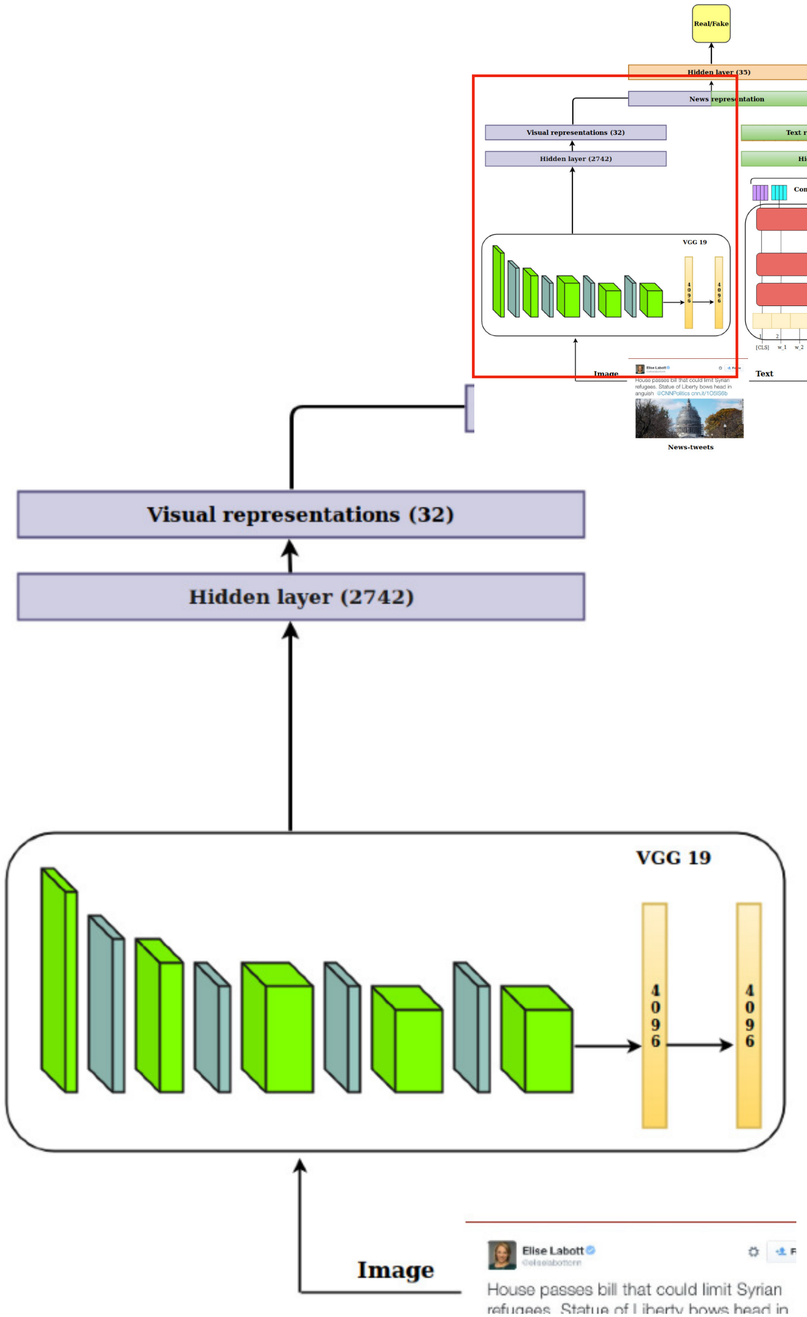
- It uses BERT to represent words and sentences in a way that best captures underlying semantic and contextual meaning.
- The features obtained from BERT-base are the desired contextual embedding of the post that are then passed through a fully-connected layer to reduce down to final dimension of length 32.



Methodology

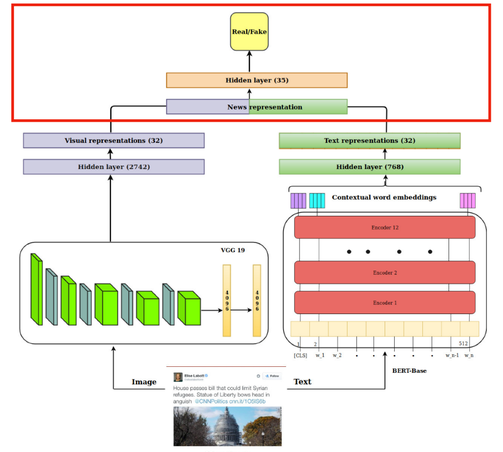
Visual Feature Extractor

- Employ the pre-trained VGG-19.
- Extract the output of the second last layer of VGG-19 and passed it through a fully connected layer to reduce down to final dimension of length 32.

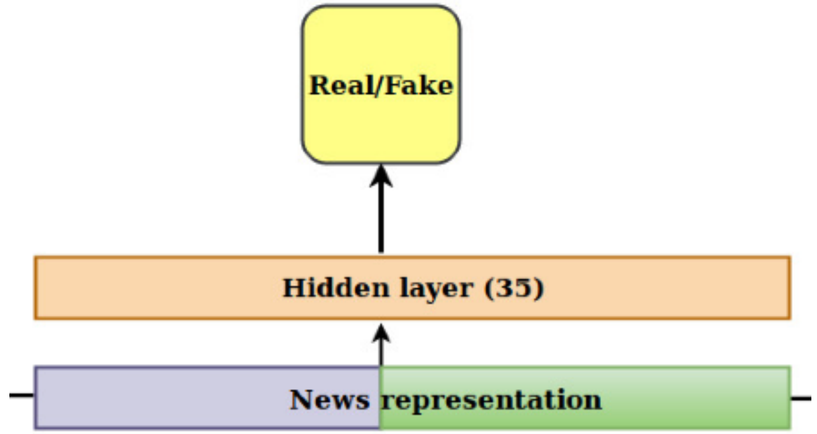


Methodology

Multimodal Fusion



- The two feature vectors obtained via different modalities are fused together using simple concatenation technique to obtain the desired news representation.
- Then passed through a fully connected neural network for fake news classification.



Experiments

Dataset & Hyperparameter

- Twitter (MediaEval-16)
- Weibo

parameters	Twitter	Weibo
BERT trainable	False	False
VGG trainable	False	False
dropout	0.4	0.4
# of hidden layers (text)	2	2
# of neurons in hidden layer (text)	768,32	768,32
# of hidden layers (image)	2	1
# of neurons in hidden layer (image)	2742,32	32
# of dense layers (concatenation)	1	1
# of neurons in dense layer (concatenate)	64	64
text length	23 words	200 char
batch size	256	256
optimizer	adam	adam
learning rate	0.0005	0.001

Experiments

Baselines - Single Modality

- Textual: uses only textual information present in posts to classify them as fake or not.
- Visual: uses only images from the posts to classify them as fake or not.

Experiments

Baselines - Multimodal Models

- VQA: Visual Question Answering aims to questions about given images. Adapted the Visual QA model which was originally designed for a multi-class classification task to binary classification.
- Neural Talk: image captioning, proposes generation of natural language sentences describing an image using a deep recurrent framework.
- att-RNN: use attention mechanisms to combine textual, visual and social context features.
- EANN: built an end-to-end model for fake news detection and event discriminator.
- MVNN: build an auto encoder-decoder model for fake news detection.

Experiments

Results

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precison	Recall	F1-Score
Twitter	textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA [27]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk [28]	0.610	0.728	0.504	0.595	.534	0.752	0.625
	att-RNN [29]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN- [20]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	EANN [20]	0.715	NA	NA	NA	NA	NA	NA
	MVAE- [21]	0.656	NA	NA	0.641	NA	NA	0.669
	MVAE [21]	0.745	**0.801**	0.719	0.758	0.689	**0.777**	**0.730**
	SpotFake	**0.7777**	0.751	**0.900**	**0.82**	**0.832**	0.606	0.701
Weibo	textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN-	0.795	0.827	0.697	0.756	0.752	0.863	0.804
	EANN	0.827	NA	NA	NA	NA	NA	NA
	MVAE-	0.743	NA	NA	NA	NA	NA	NA
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	**0.837**
	SpotFake	**0.8923**	**0.902**	**0.964**	**0.932**	**0.847**	0.656	0.739

- Reporting the performance comparison of SpotFake with EANN & MVAE on accuracy %.
- EANN- / MVAE- is when fake news classifier is trained standalone (w/o secondary task).
  - EANN: event discriminator that removes the event-specific features and keep shared features among events.
  - MVAE: discover the correlations across the modalities to improve shared representations.

Experiments

Results

Dataset	Model	Accuracy	Fake News			Real News		
			Precision	Recall	F1-Score	Precison	Recall	F1-Score
Twitter	textual	0.526	0.586	0.553	0.569	0.469	0.526	0.496
	visual	0.596	0.695	0.518	0.593	0.524	0.7	0.599
	VQA [27]	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk [28]	0.610	0.728	0.504	0.595	.534	0.752	0.625
	att-RNN [29]	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN- [20]	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	EANN [20]	0.715	NA	NA	NA	NA	NA	NA
	MVAE- [21]	0.656	NA	NA	0.641	NA	NA	0.669
	MVAE [21]	0.745	**0.801**	0.719	0.758	0.689	**0.777**	**0.730**
	SpotFake	**0.7777**	0.751	**0.900**	**0.82**	**0.832**	0.606	0.701
Weibo	textual	0.643	0.662	0.578	0.617	0.609	0.685	0.647
	visual	0.608	0.610	0.605	0.607	0.607	0.611	0.609
	VQA	0.736	0.797	0.634	0.706	0.695	0.838	0.760
	Neural Talk	0.726	0.794	0.713	0.692	0.684	0.840	0.754
	att-RNN	0.772	0.797	0.713	0.692	0.684	0.840	0.754
	EANN-	0.795	0.827	0.697	0.756	0.752	0.863	0.804
	EANN	0.827	NA	NA	NA	NA	NA	NA
	MVAE-	0.743	NA	NA	NA	NA	NA	NA
	MVAE	0.824	0.854	0.769	0.809	0.802	**0.875**	**0.837**
	SpotFake	**0.8923**	**0.902**	**0.964**	**0.932**	**0.847**	0.656	0.739

- Though SpotFake is a standalone fake news classifier, still outperform both configurations of EANN and MVAE by large margins.

	Accuracy	
Model	**Twitter**	**Weibo**
EANN- [20]	64.8 (12.97)	79.5 (9.73)
EANN [20]	71.5 (6.27)	82.7 (6.53)
SpotFake	**77.77**	**89.23**

	Accuracy	
Model	**Twitter**	**Weibo**
MVAE- [21]	65.6 (12.17)	74.3 (14.93)
MVAE [21]	74.5 (3.27)	82.4 (6.83)
SpotFake	**77.77**	**89.23**

Conclusions

- Previous literature has attacked the problem of detecting fake news from different angles like NLP, KG, CV and user profiling.
  - It has been shown that for consistent results → a multimodal method is required.
- SpotFake uses language transformer model and pre-trained ImageNet models for extraction and classifies using fully connected layer.
- It outperforms the baselines by a margin of 6% accuracy on an average.
- Still room for improvement on longer length articles and more complex fusion techniques to understand how different modalities play a role in fake news detection.

The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

SpotFake+: A Multimodal Framework for Fake News Detection via Transfer Learning (Student Abstract)

Shivangi Singhal,[1] Anubha Kabra,[2*] Mohit Sharma,[3*]
Rajiv Ratn Shah,[4] Tanmoy Chakraborty,[5] Ponnurangam Kumaraguru[6]
[1,3,4,5,6]IIIT-Delhi, India, [2]DTU, India
(shivangis[1], rajivratn[4], tanmoy[5], pk[6])@iiitd.ac.in, {anubhakabradtu, mohit.sharma.cs29}@gmail.com[2,3]

AAAI'20

220126 Chia-Chun Ho

Introduction

Datasets changed

- In this paper, the authors consider the FakeNewsNet repository for multimodal fake news detection.
- In contrast to existing datasets (MediaEval-16, Weibo) in this space, FakeNewsNet consists of full length articles rather than short claims or news in the form of tweets.
- FakeNewsNet also contains image associated with each article.
- Thus the authors believe that it's more representative of a news article.

Introduction

Previous studies

- Have used various machine learning techniques (like SVM, Naive Bayes, Logistic Regression) and deep learning models (CNN, LSTM, Attention).
- But they fail to perform well due to:
  - Lacked the contextual information present in the text.
  - Do not capture the features from the image modality that may seek to emphasize certain fact.

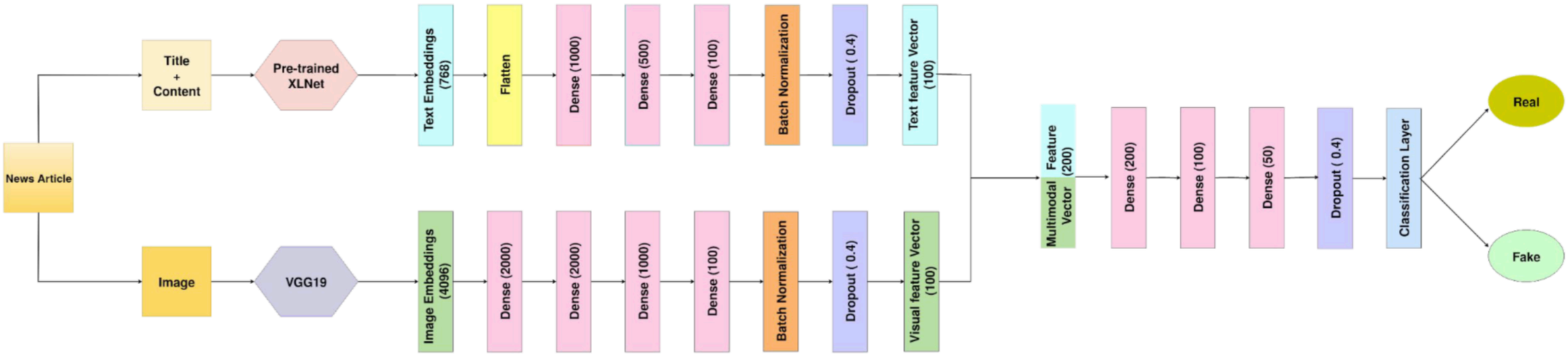
Introduction

SpotFake+

- To overcome the above mentioned challenges, proposed SpotFake+.
  - An advanced version of existing multimodal fake news detection system SpotFake.
  - Leverage transfer learning to capture semantic and contextual information from the news articles and its associated images.
  - First work that performs a multimodal approach for fake news detection on a dataset that consists of full length articles.

Methodology

Framework overview



Experiments

Results

Dataset	Politifact	Gossipcop
Real	624 (321)	16817 (10259)
Fake	432 (164)	5323 (2581)

# of samples in the FakeNewsNet repository.
The values in the brackets indicate samples fit to use after data pre-processing.

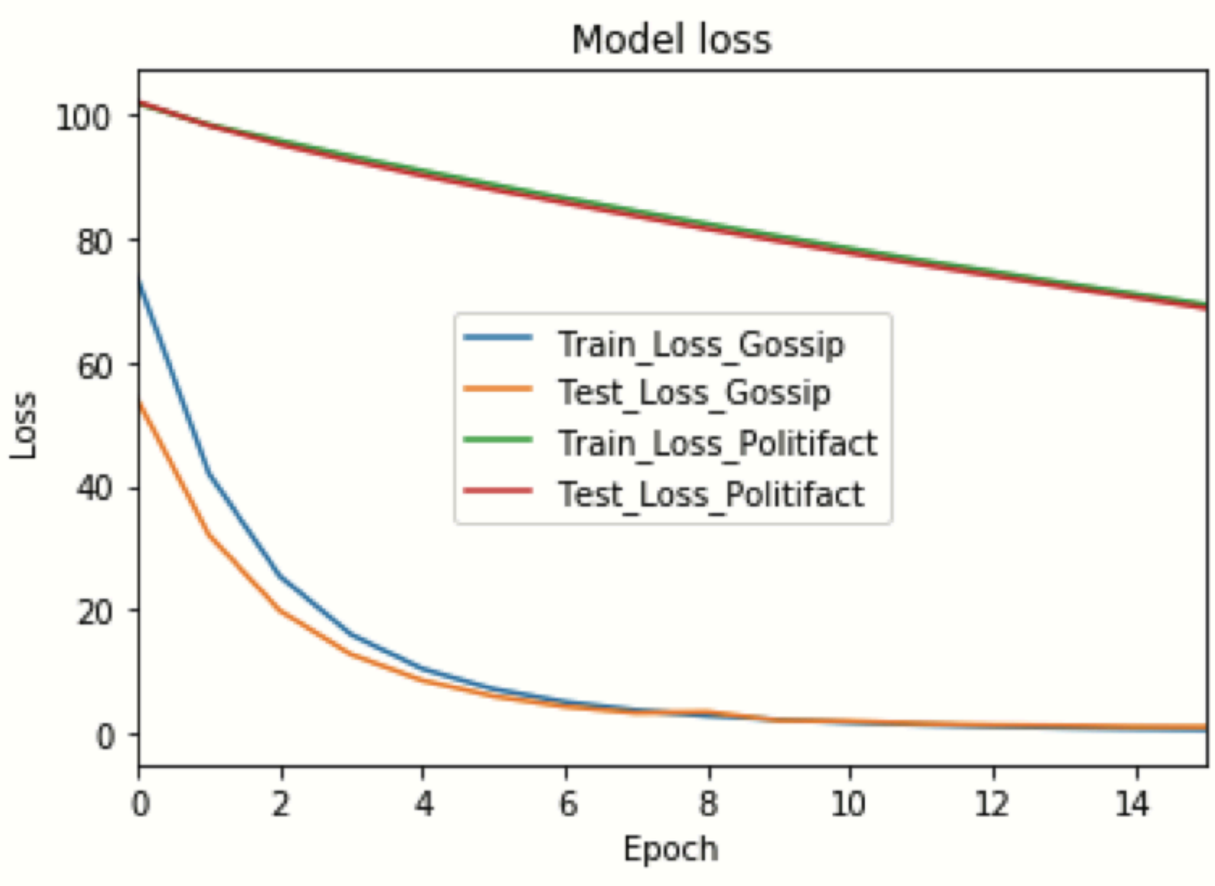
Modality	Models	Politifact	Gossipcop
Text	SVM	0.58	0.497
	Logistic Regression	0.642	0.648
	Naive Bayes	0.617	0.624
	CNN	0.629	0.723
	SAF (Social Article Fusion)	0.691	0.689
	XLNet + dense layer	**0.74**	**0.836**
	XLNet+ CNN	0.721	0.84
	XLNet + LSTM	0.721	0.807
Image	VGG19	0.654	0.80
Multimodal (Text+Image)	EANN (Wang et al. 2018)	0.74	0.86
	MVAE (Khattar et al. 2019)	0.673	0.775
	SpotFake (Singhal et al. 2019)	0.721	0.807
	SpotFake+ (XLNet + dense + VGG19)	**0.846**	**0.856**

Experiments

Loss function graph

Dataset	Politifact	Gossipcop
Real	624 (321)	16817 (10259)
Fake	432 (164)	5323 (2581)

# of samples in the FakeNewsNet repository.
The values in the brackets indicate samples fit to use after data pre-processing.



Conclusion & Future Works

- Present SpotFake+, advance version of SpotFake.
- Proposed architecture uses transfer learning to capture the textual and visual features within an article.
- Experiments performed in this paper further reveal the potential of multimodal features for the problem of fake news detection.
- The work can further be expanded to incorporate meta level feature modalities.

Comments
of SpotFake / SpotFake+

- Just simple textual and visual representation concatenations.
- SpotFake+ has mentioned used transfer learning but didn't see detail informations.
- SpotFake+ use XLNet for textual extraction instead of BERT.
- In FakeNewsNet, observed that EANN is perform well than SpotFake.
  - In GossipCop, EANN even outperform all methods.
- May can enhance by additional metadata information.