MVAE: Multimodal Variational Autoencoder for Fake News Detection

Dhruv Khattar
International Institute of Information Technology
Hyderabad, India
dhruv.khattar@research.iiit.ac.in

Jaipal Singh Goud
International Institute of Information Technology
Hyderabad, India
jaipal.singh@research.iiit.ac.in

Manish Gupta*
International Institute of Information Technology
Hyderabad, India
manish.gupta@iiit.ac.in

Vasudeva Varma
International Institute of Information Technology
Hyderabad, India
vv@iiit.ac.in

WWW'19

220118 Chia-Chun Ho

Outline

Introduction

Methodology

Experiments

Conclusions

Comments

Introduction

Fake news examples

- Each tweet has certain textual content and an image associated with it.
- For the tweet on the left, both the image and text indicate that it's probably a fake news.
- In the tweet on the right, the image doesn't add substantial information but the text indicates that it may be a fake news.

Text: **Elephant carved out of solid rock. Magnificent!**

Text: **New species of fish found in Arkansas.**

Text: **Woman, 36, gives birth to 14 children from 14 different fathers**

Figure 1: Fake News examples from the Twitter dataset

Introduction

Fake news examples

- In the tweet in the middle, it's difficult to reach a conclusion from the text, but the morphed image suggests that it's possibly a fake news.
- This example reflects the hypothesis that pairs of visual and textual information can give better insights into fake news detection.

Text: **Elephant carved out of solid rock. Magnificent!**

Text: **New species of fish found in Arkansas.**

Text: **Woman, 36, gives birth to 14 children from 14 different fathers**

Figure 1: Fake News examples from the Twitter dataset

Introduction

On conceptual level

- Detecting fake news has undergone a variety of labels from misinformation to rumor.
- The target of study is detection of news content that is fabricated and can be verified to be false.
- Rumor and fake news detection techniques range from traditional learning methods to deep learning models.

Introduction

Related works

- Initial approaches tried to detect fake news using only linguistic features extracted from the text content of news stories.
- Some work explored the possibility of representing tweets with deep neural networks by capturing temporal-linguistic features.
- Recent works in deep learning to detect fake news have shown performance improvements over traditional methods due to their enhanced ability to extract relevant features.

Introduction

Related works

- att-RNN: combine visual, textual and social context features, using an attention mechanism to make predictions about fake news.
- EANN: use an additional event discriminator to learn common features shared among all the event-specific features, and claim that they can handle novel and newly emerged events better.
- A shortcoming of the existing models is that they don't have any explicit objective function to discover correlations across the modalities.

Introduction

MVAE

- Inspired by the idea of auto encoders tries to learn shared representations in a multimodal setting.
- Variational Autoencoder (VAE)s can learn probabilistic latent variable models by optimizing a bound on the marginal likelihood of the observed data.
- Overcoming the limitations of the current models, proposed a multimodal variational autoencoder capable of learning shared (visual+textual) representations, trained to discover correlations across modalities in tweets.
- The VAE is then coupled with a classifier to detect fake news.

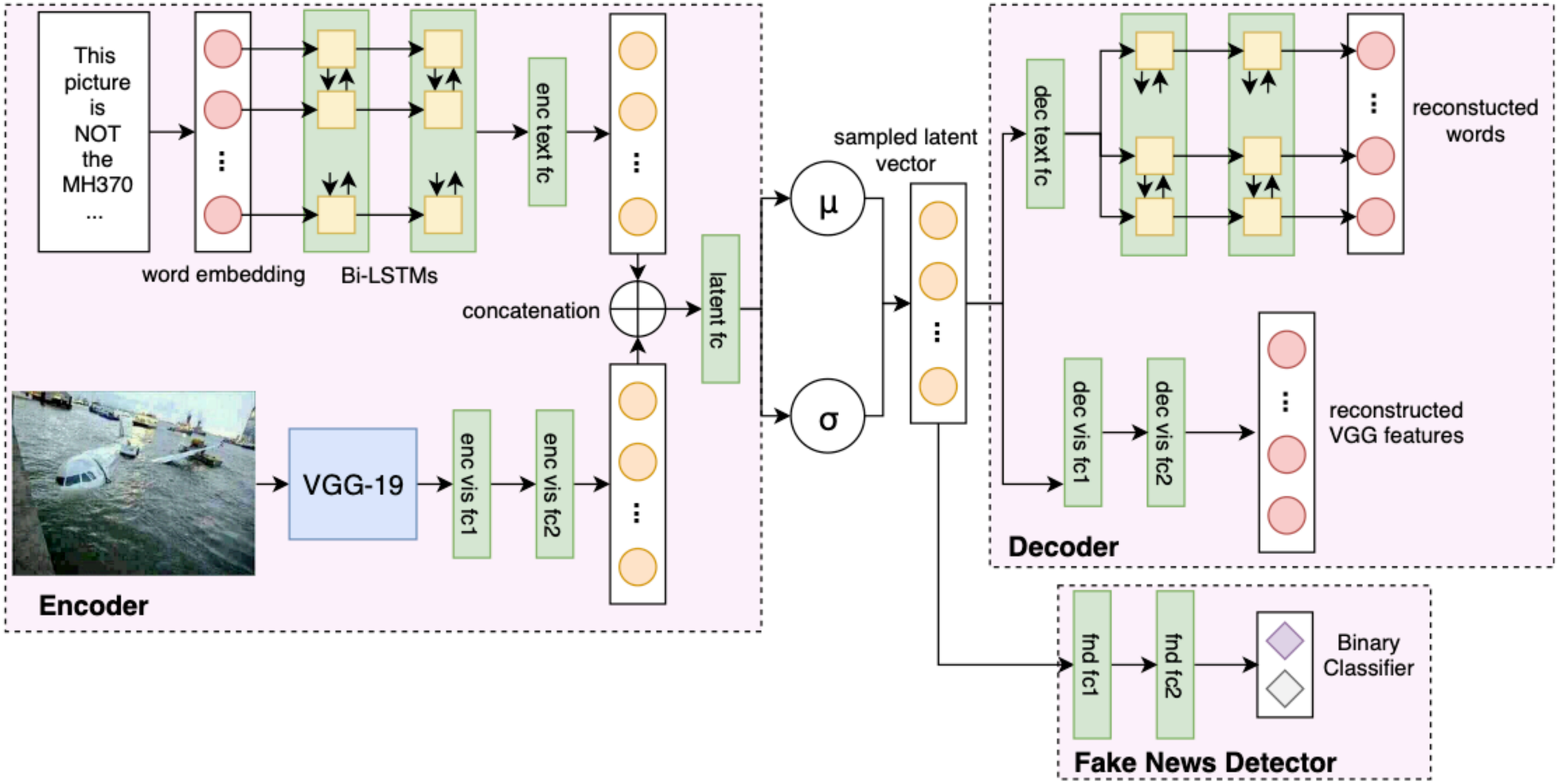
Introduction

Contributions

- Propose a novel approach for classifying social media posts using only the content of the posts using only the post, i.e., the text and the attached image.
- The proposed MVAE model uses a Multimodal Variational Autoencoder trained jointly with a Fake News Detector to detect if a post is fake or not.
- Extensively evaluate the performance of proposed model on two real-world datasets. The results reveal that proposed model learns better multimodal features and outperforms the SOTA multimodal fake news detection models.
- Show that proposed model is able to discover correlations across the modalities and thus come up with better multimodal shared representations.

Methodology

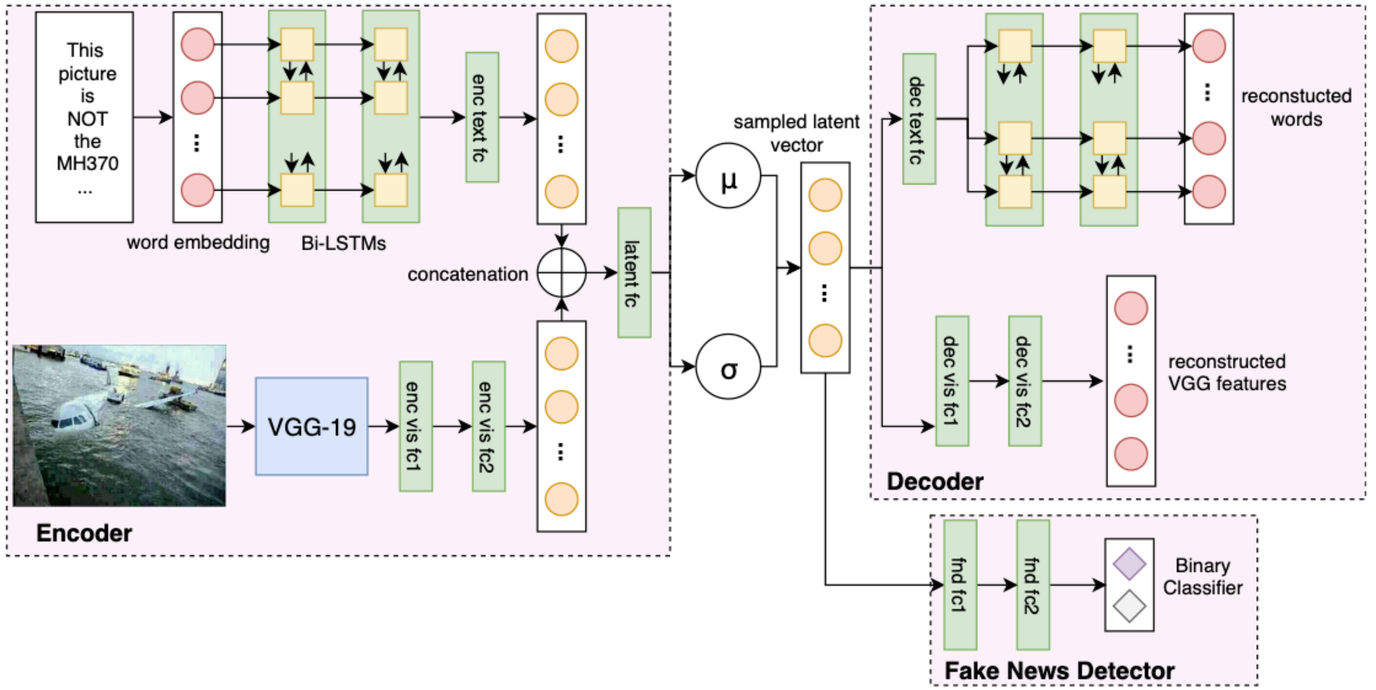
Framework overview



Methodology

Overview

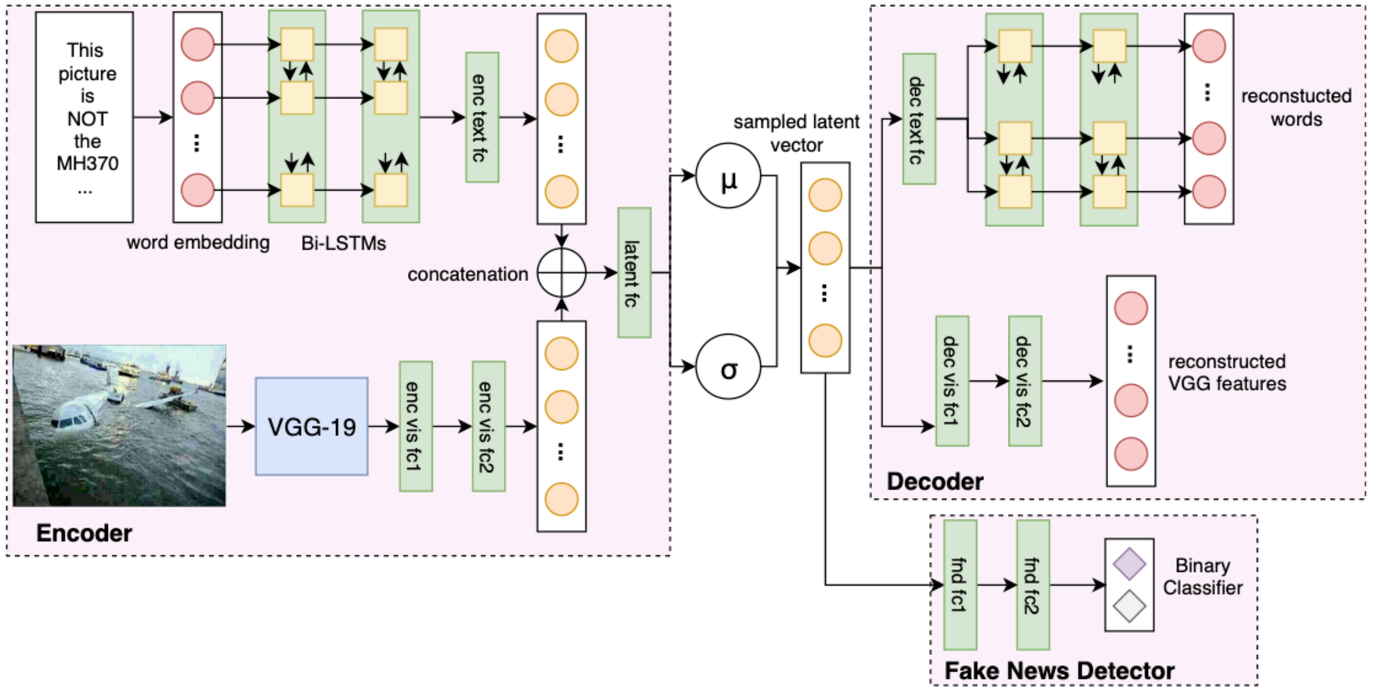
- Address the problem of fake news detection.
- The basic idea behind MVAE is to learn a unified representation of both the modalities of a tweet's content.



Methodology

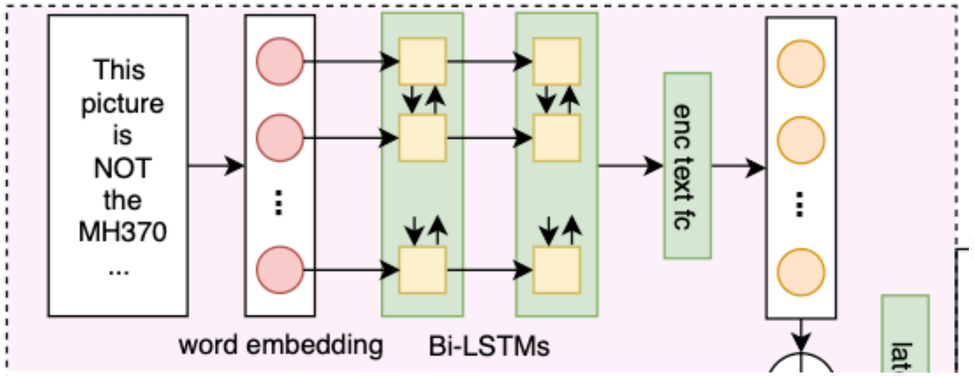
Overview

- Has three components:
  - Encoder: encodes the information from text and image into a latent vector.
  - Decoder: reconstructs back the original image and text from the latent vector.
  - Fake News Detector: uses the learned shared representation to predict if a news is fake or not.



Methodology

Textual Encoder

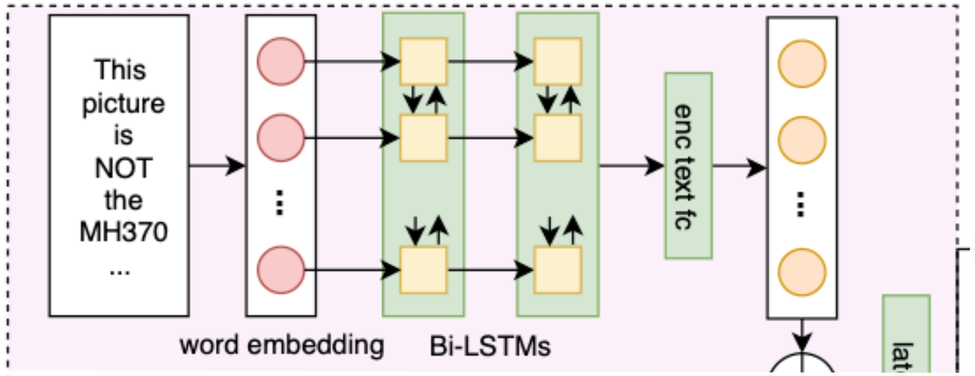


- Input: $T = [T_1, T_2, \cdots T_n]$
- n: # of words in the text
- $T_i \in T$: word embedding vector obtained from pre-trained model

- Employ stacked Bi-LSTM units to extract textual features.
- The final hidden state of LSTM units to extract textual features.
- The final hidden state of LSTM is obtained by concatenating the forward and backward states.

Methodology

Textual Encoder



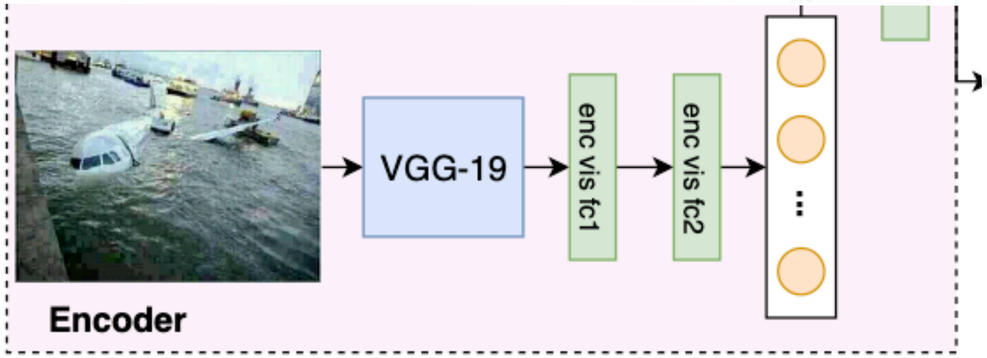
- Input: $T = [T_1, T_2, \cdots T_n]$
- n: # of words in the text
- $T_i \in T$: word embedding vector obtained from pre-trained model

- Finally, pass the LSTM output through a fully connected layer to get the textual features.
- $R_T = \phi(W_{tf} R_{lstm})$
  - R_{lstm}: output of the LSTM
  - W_{tf}: weight matrix of fully connected layer
  - ϕ: activation function

Methodology

Visual Encoder

- Input: V image attached with the post

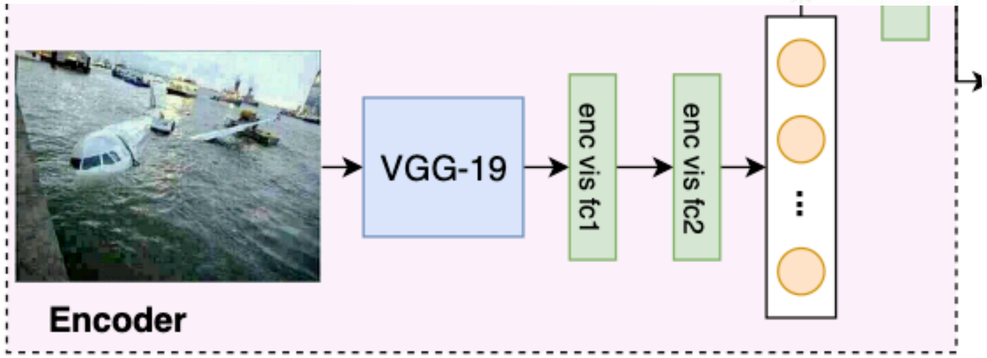


- Employ the pre-trained network of VGG-19 architecture trained over the ImageNet database.
- Freeze the parameters of the VGG network to avoid parameter explosion.

Methodology

Visual Encoder

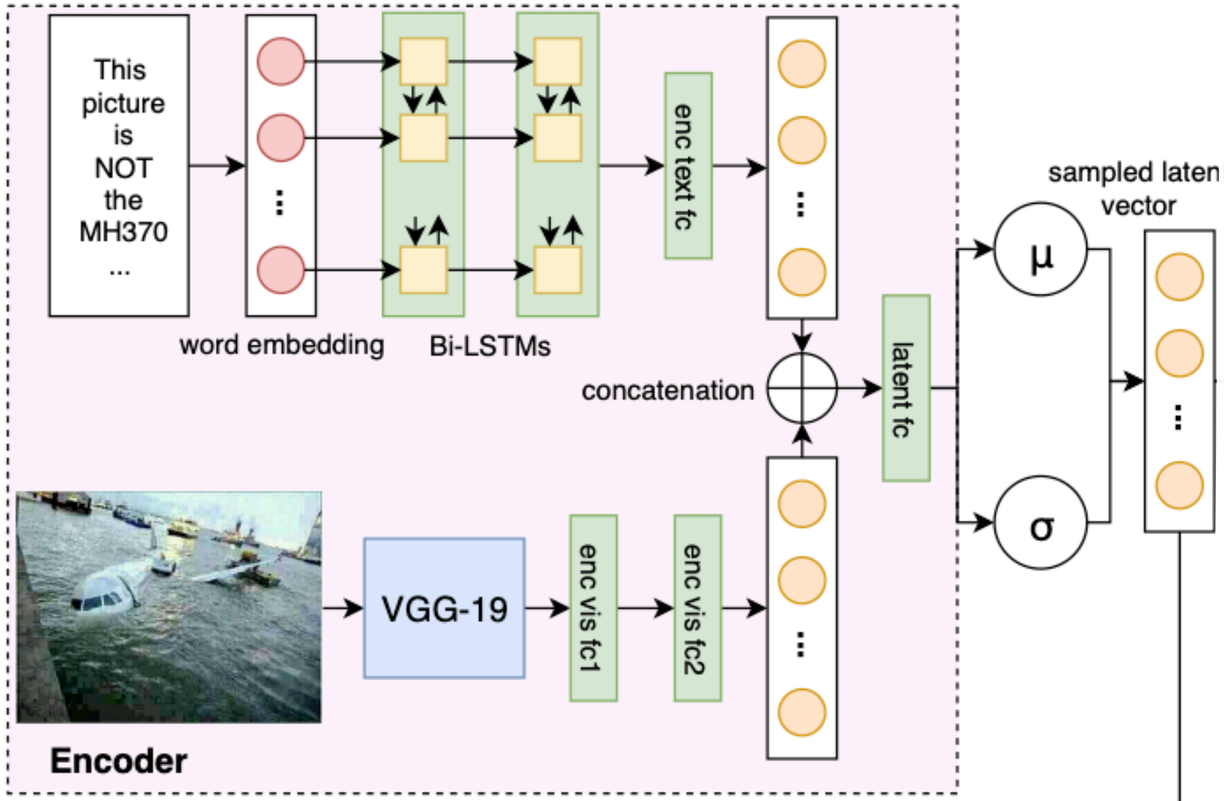
- Input: V image attached with the post



- Finally, pass the VGG output through multiple fully connected layers to get the same sized representation of the image as that of text.
- $R_V = \phi(W_{vf} R_{vgg})$
  - R_{vgg}: representation obtained from VGG-19
  - W_{vf}: weight matrix of the fully connected layer

Methodology

Encoder

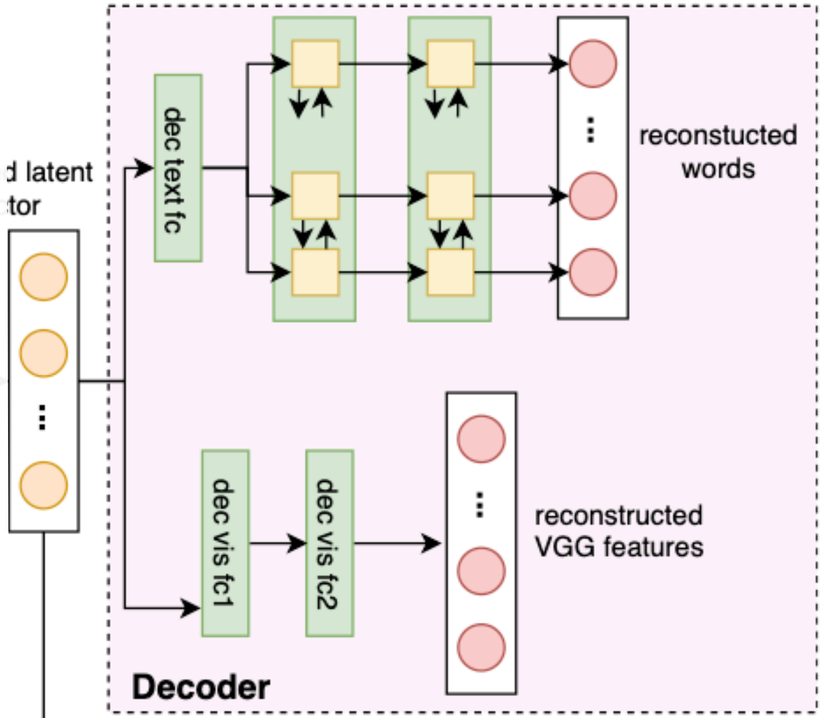


- Denote encoder as:
- $R_m = G_{enc}(m, \theta_{enc})$
  - θ_{enc}: all parameters to be learned in encoder
  - m: multimedia post

Methodology

Decoder

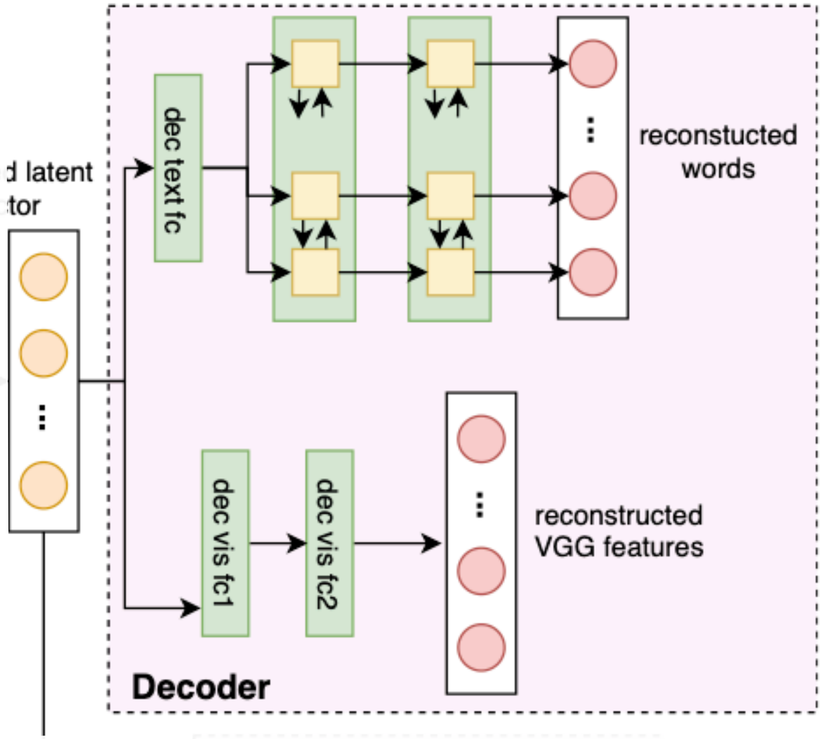
- Textual decoder takes as input the multimodal representation and reconstructs the words in the text.
- The multimodal representation is pass through a fully connected layer to create inputs for the Bi-LSTM.
- Finally, pass the LSTM outputs through a time distributed fully connected layer with softmax activation to get the probability of each word in that time step.



Methodology

Decoder

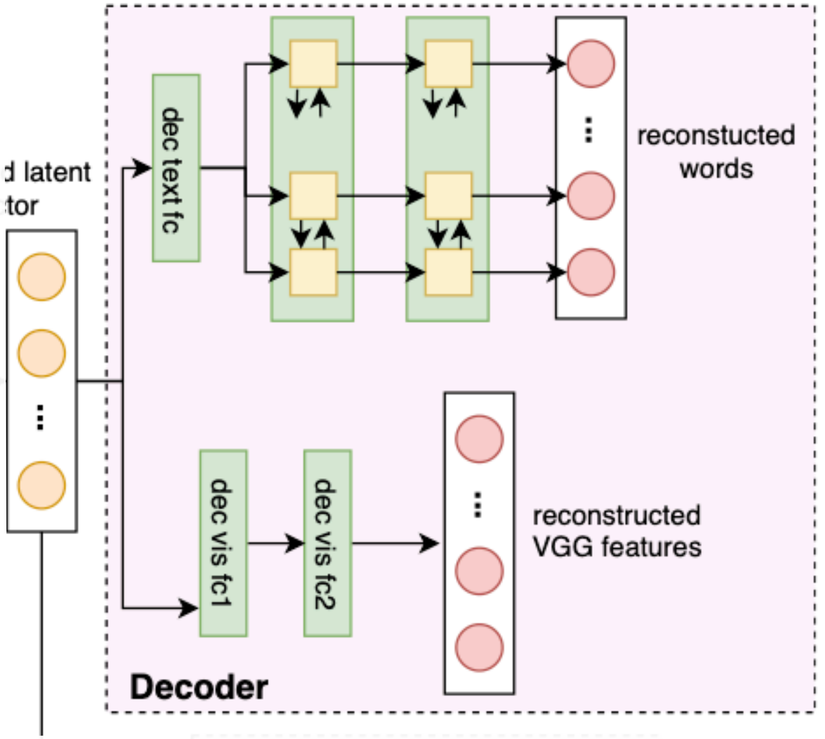
- Visual decoder is to reconstruct the VGG-19 features from the multimodal representation.
- Just the reverse of the visual encoder's corresponding sub-network.
- The multimodal representation is passed through multiple fully connected layers to reconstruct the VGG-19 features.



Methodology

Decoder

- Denote the decoder as $G_{dec}(R_m, \theta_{dec})$
- $(\hat{t}_m, \hat{r}_{vgg_m}) = G_{dec}(R_m, \theta_{dec})$
  - θ_{dec}: all the parameters in the decoder
  - $\hat{t}_m$: matrix of probability of every word for each position in the text
  - $\hat{r}_{vgg_m}$: reconstructed VGG-19 features of the image



Methodology

Optimizing

- VAE models are trained by optimizing the sum of the reconstruction loss and the KL divergence loss.
- Therefore, employ categorical cross-entropy loss for reconstruction of the text mean squared error for reconstruction of image features.
- Minimizing the KL divergence here means optimizing the probability distribution parameters (μ, σ) to closely resemble that of the target distribution.

Methodology

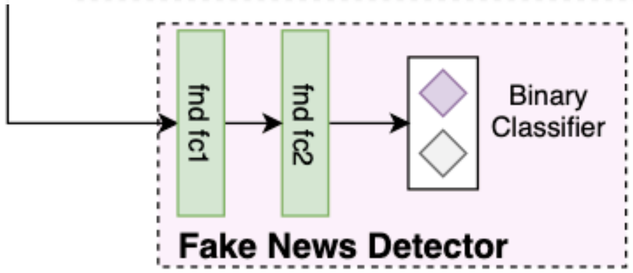
Optimizing

- $$\mathscr{L}_{rec_{vgg}} = \mathbb{E}_{m \sim M}[\frac{1}{n_v} \sum_{i=1}^{n_v} (\hat{r}_{vgg_m}^{(i)} - r_{vgg_m}^{(i)})^2]$$
- $$\mathscr{L}_{rec_t} = -\mathbb{E}_{m \sim M}[\sum_{i=1}^{n_t} \sum_{c=1}^{C} 1_{c=t_m^{(i)}} \log \hat{t}_m^{(i)}]$$
- $$\mathscr{L}_{kl} = \frac{1}{2} \sum_{i=1}^{n_m} (\mu_i^2 + \sigma_i^2 - \log(\sigma_i) - 1)$$
- $$(\theta_{enc}^*, \theta_{dec}^*) = \underset{\theta_{enc}, \theta_{dec}}{\operatorname{argmin}} (\mathscr{L}_{rec_{vgg}} + \mathscr{L}_{rec_t} + \mathscr{L}_{kl})$$

Methodology

Fake news detector

- Take multimodal representation as input and aims to classify the post as fake or not.
- It consists of multiple fully connected layers with corresponding activation functions.
- $\hat{y}_m = G_{fnd}(R_m, \theta_{fnd})$
- $\mathscr{L}_{fnd}(\theta_{enc}, \theta_{fnd}) = -\mathbb{E}_{(m,y)\sim(M,Y)}[y \log(\hat{y}_m) + (1-y)\log(1-\hat{y}_m)]$
- $(\theta^*_{enc}, \theta^*_{fnd}) = \underset{\theta_{enc},\theta_{fnd}}{\operatorname{argmin}} \mathscr{L}_{fnd}$



Methodology

Putting it all together

- The output of the encoder is fed to the decoder as well as the fake news detector.
- The decoder aims to reconstruct the data, while the fake news detector aims to classify the post as fake news or not.
- Jointly train the VAE and the fake news detector.
  - $\mathscr{L}_{final}(\theta_{enc}, \theta_{dec}, \theta_{fnd}) = \lambda_v \mathscr{L}_{rec_{vgg}} + \lambda_t \mathscr{L}_{rec_t} + \lambda_k \mathscr{L}_{kl} + \lambda_f \mathscr{L}_{fnd}$
  - $(\theta^*_{enc}, \theta^*_{dec}, \theta^*_{fnd}) = \underset{\theta_{enc},\theta_{dec},\theta_{fnd}}{\operatorname{argmin}} \mathscr{L}_{final}(\theta_{enc}, \theta_{dec}, \theta_{fnd})$

Experiments

Dataset

- Twitter (MediaEval-16)
  - 9000 fake news tweets + 6000 real news tweets + 2000 test tweets = 17000 unique
- Weibo
  - 4:1 training : testing

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: w/o event discriminator for fair comparison

Experiments

Results:

On Twitter Dataset

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
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	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	**0.810**	0.498	0.617	0.584	0.759	0.660
	MVAE	**0.745**	0.801	**0.719**	**0.758**	**0.689**	**0.777**	**0.730**
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	0.854	0.656	0.742	0.72	**0.889**	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	**0.824**	**0.854**	**0.769**	**0.809**	**0.802**	0.875	**0.837**

- Proposed method performs much better than the baselines.
- Visual model performs better than the textual model.
  - Image features learnt with the help of VGG-19 have more shareable patterns to classify news as compared to textual features.
  - Still works than the multimodal models.

Experiments

Results:

On Twitter Dataset

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
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	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	**0.810**	0.498	0.617	0.584	0.759	0.660
	MVAE	**0.745**	0.801	**0.719**	**0.758**	**0.689**	**0.777**	**0.730**
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	0.854	0.656	0.742	0.72	**0.889**	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	**0.824**	**0.854**	**0.769**	**0.809**	**0.802**	0.875	**0.837**

- att-RNN beats EANN which tells us that the attention mechanism can help in improving the performance of the model by considering the parts of the image which are related to the text.

- MVAE outperforms the baseline models by huge margin and increases the accuracy 66.4 → 74.5 and increases the F1 score 66 → 73.

Experiments

Results:

On Weibo Dataset

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F_1	Precision	Recall	F_1
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	0.631	0.765	0.509	0.611	0.55	0.794	0.65
	Neural Talk	0.610	0.728	0.504	0.595	0.534	0.752	0.625
	att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	**0.810**	0.498	0.617	0.584	0.759	0.660
	MVAE	**0.745**	0.801	**0.719**	**0.758**	**0.689**	**0.777**	**0.730**
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	0.854	0.656	0.742	0.72	**0.889**	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	**0.824**	**0.854**	**0.769**	**0.809**	**0.802**	0.875	**0.837**

- See similar trends in the results.
- The textual model beats the visual model in single modality models.
- EANN & att-RNN which were proposed for this task perform better than Neural Talk and VQA.
- MVAE outperforms all the baselines and boosts the performance, this validates the effectiveness of proposed method MVAE in detecting fake news on social media.

Conclusions

- Propose a multimodal variational autoencoder that learns shared (visual + textual) representations to aid fake news detection.
- MVAE is trained by jointly learning the encoder, decoder and the fake news detector.
- The MVAE model outperforms the current SOTA architectures. (Now out of date)
- Plan to extend MVAE using tweet propagation data and user characteristics.

Comments
of MVAE

- Use concept of auto encoder.
- Discover the difference between textual and visual feature representation.
- EANN in this paper w/o discriminator lead performance down.