EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao.
In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18).
Association for Computing Machinery, New York, NY, USA, 849–857.
DOI:https://doi.org/10.1145/3219819.3219903

20210318 何嘉峻

Outline

Introduction

Related Work

Methodology

Experiments

Conclusions and Contribution

Comments

Introduction.

Fake News Detection existing methods

- Traditional learning and deep learning based models
  - Existing deep learning models have achieved performance improvement over traditional ones due to the ability of feature extraction.
- BUT still can't handle the situation like detecting fake news **on newly emerged and time-critical events.**
  - New event data leads to unsatisfactory performance of existing models.

Introduction..

Handle the newly emerged event

- Existing models tend to capture lots of event-specific feature (which not shared among different events).
  - Event-specific features being able to help classify the post on verified events
  - But hurt the detection with regard to newly emerged events
- **Learning the shared features among all the events** would help to detect the fake news from unverified post.

Introduction...

Learning the shared features among all the events

- Design an effective model to remove the nontransferable event-specific feature and preserve the shared features among all the events.
- Identify the event-specific features first
  - Posts on different events have their own unique of specific feature that are not sharable.
  - Can be detected by measuring the difference among posts corresponding to different events.

Introduction....

Identify the event-specific features

- It's equivalent to measuring the difference among learned features on different events.

1. Feature representation of posts are high-dimensional
    - Simple metrics like squared error may not be able to estimate.
2. Feature representation keep changing during the training stage.
    - Required the proposed measurement mechanism to capture the changes of feature representations and consistently provide the accurate measurement.

Introduction.....

Event Adversarial Neural Networks (EANN)

- Inspired by adversarial network, use the event discriminator to predict the event auxiliary labels during training stage, and the corresponding loss can be used to estimate the dissimilarities of feature representations among different events.
  - The larger the loss, the lower the dissimilarities.
- fake news takes advantage of multimedia content to mislead readers and gets spread
  - model needs to handle the multi-modal inputs

Introduction.....

Event Adversarial Neural Networks (EANN)

- The proposed model EANN consists of three main components:
  - **multi-modal feature extractor**
    - employ CNN to automatically extract features from both textual and visual
  - **fake news detector**
  - **event discriminator**
    - multi-modal feature extractor tries to fool the event discriminator to learn the event invariant representations

Related Work.

Fake news detection

- **Single modality based**
  - Textual features
    - Traditional machine learning based
    - Deep learning model
  - Visual features
  - Social context features
- **Multi-modal**

Related Work..

Fake news detection

- Single modality based
  - Textual features
    - **Traditional machine learning based**
      - statistical or semantic features highly dependent on specific events and corresponding domain knowledge
    - **Deep learning model**
      - RNN to learn the representations of posts in a time series as textual features

Related Work...

Fake news detection

- Single modality based
  - **Visual features**
    - very limited studies are conducted on verifying the credibility of multimedia content on social media
    - The basic features of attached images in the posts are explored in the works.
      - these features are still hand-crafted and can hardly represent complex distributions of visual contents

Related Work....

Fake news detection

- Single modality based
  - **Social context features**
    - represent the user engagements of news on social media such as the number of followers, hash-tag(#) and retweets
    - aim to capture propagation patterns such as graph structure of the message propagation
      - very noisy, unstructured and labor intensive to collect
      - cannot provide sufficient information for newly emerged events

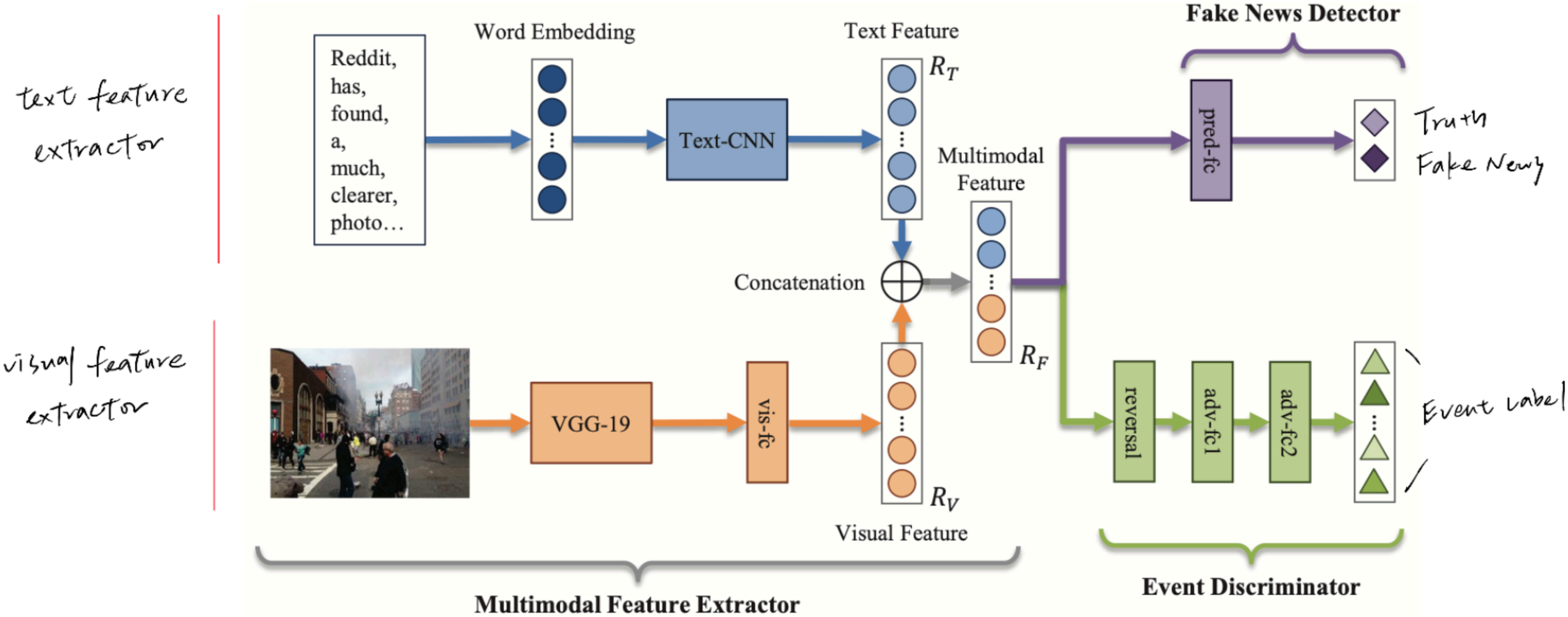
Related Work.....

Fake news detection

- **Multi-modal**
  - propose a deep learning based fake news detection model, which extracts the multi-modal and social context features and **fuses them by attention mechanism**.
  - However, the multi-modal feature representations are
    - still highly dependent on specific events in the dataset
    - cannot generalize very well to identify fake news on new coming events

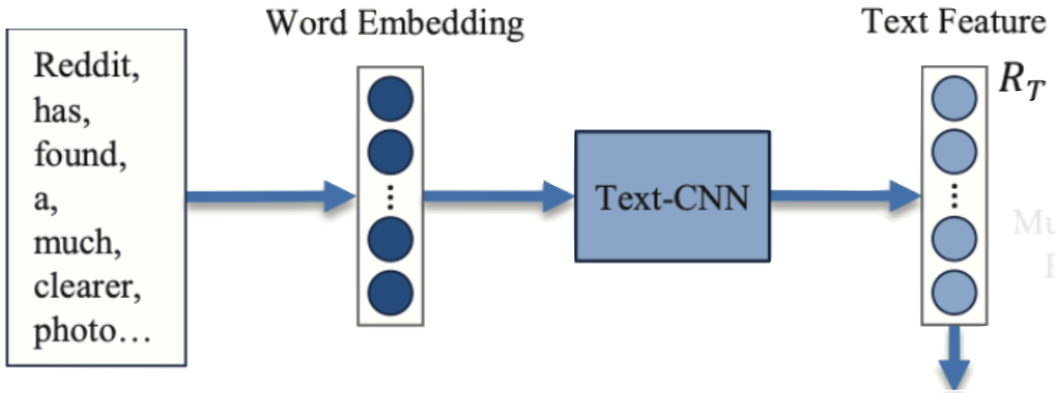
Methodology.

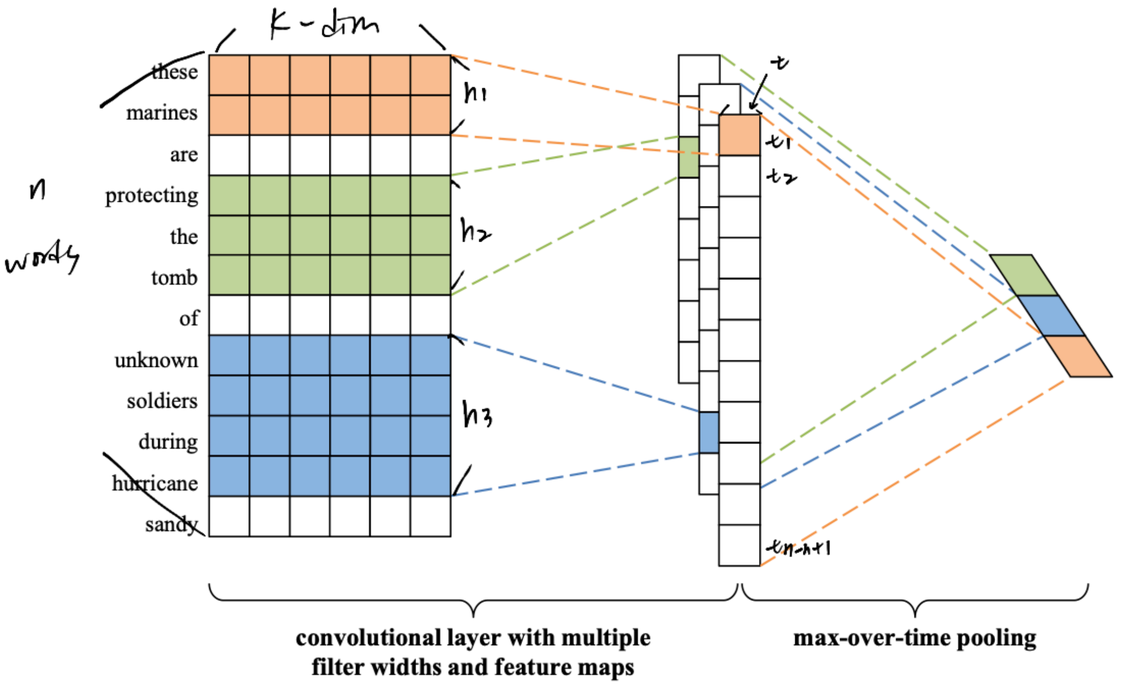
Model Overview



Methodology..

Textual Feature Extractor

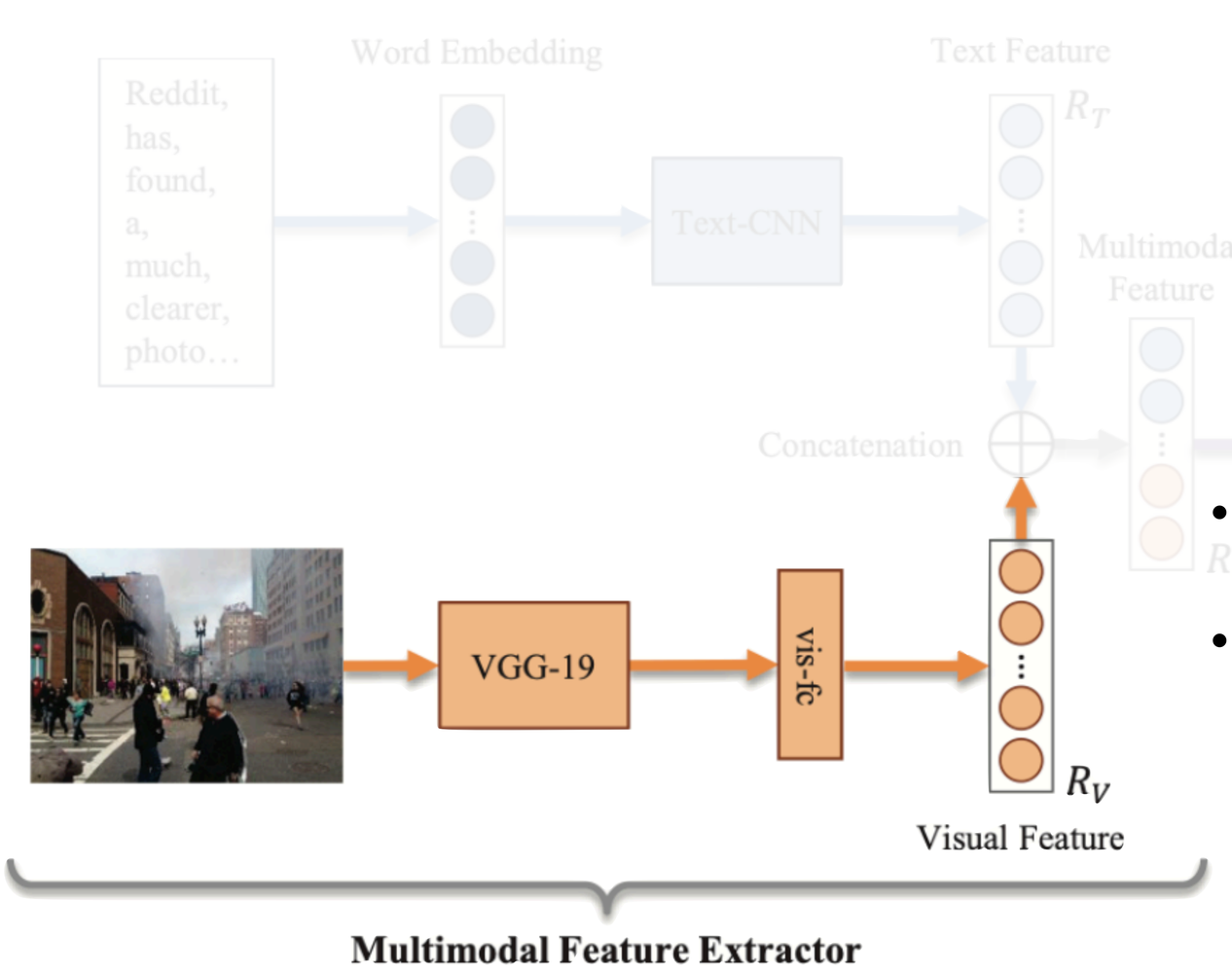




- Employ Text-CNN as text feature extractor
- n words sentence: $T_{1:n} = T_1 \oplus T_2 \cdots \oplus T_n$
- Convolutional filter with window size h: $t_i = \sigma\left(W_c \cdot T_{i:i+h-1}\right)$
- Get feature vector of sentence: $t = [t_1, t_2, \cdots, t_{n-h+1}]$
- Following the max-pooling operations, a fully connected layer to ensure the final representation ($R_T \in \mathbb{R}^p$) has the same dimension p with visual representation: $R_T = \sigma(W_{tf} \cdot R_{T_c})$

Methodology...

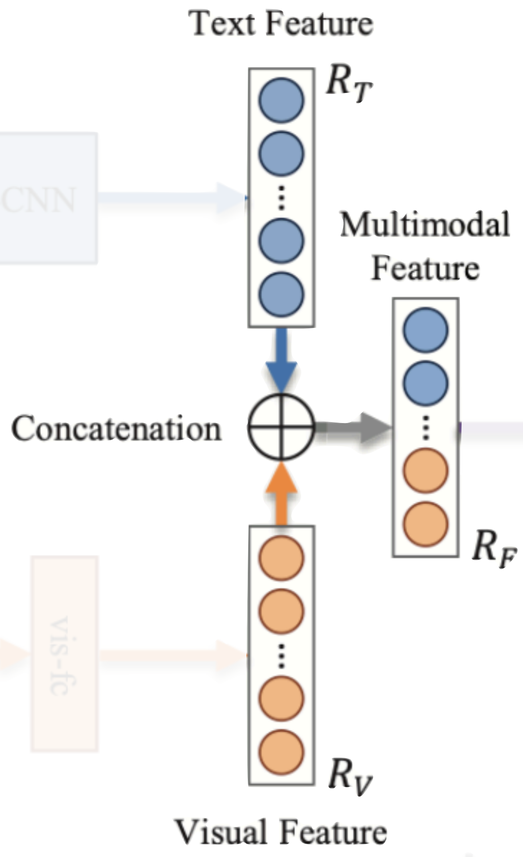
Visual Feature Extractor



- Employ VGG19 as text feature extractor
- Add a fully connected layer to adjust the dimension of final representation to p.
- $R_V = \sigma(W_{vf} \cdot R_{vgg})$

Methodology....

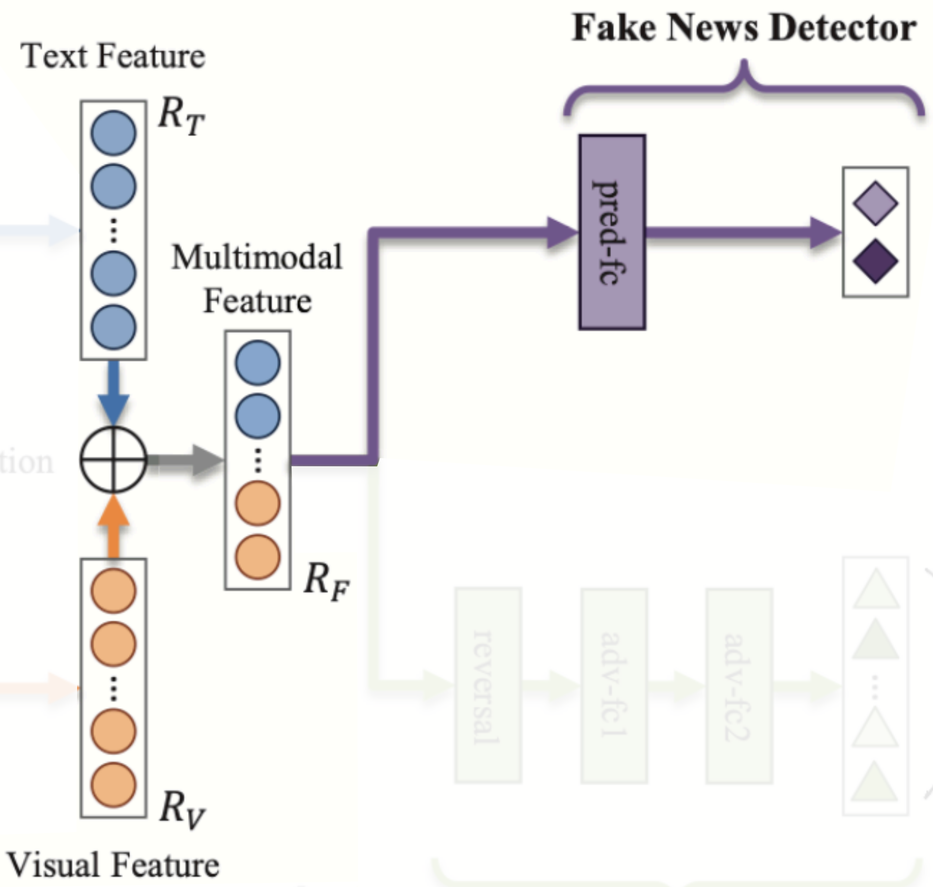
Multi-model Feature Extractor



- Concatenated to form the multi-model feature representation denoted as
  - $R_F = R_T \oplus R_R \in \mathbb{R}^{2p}$
- Denote the multi-model feature extractor
  - $G_f(M; \theta_f)$
  - M: a set of textual and visual posts
  - θ_f: parameter to be learned

Methodology.....

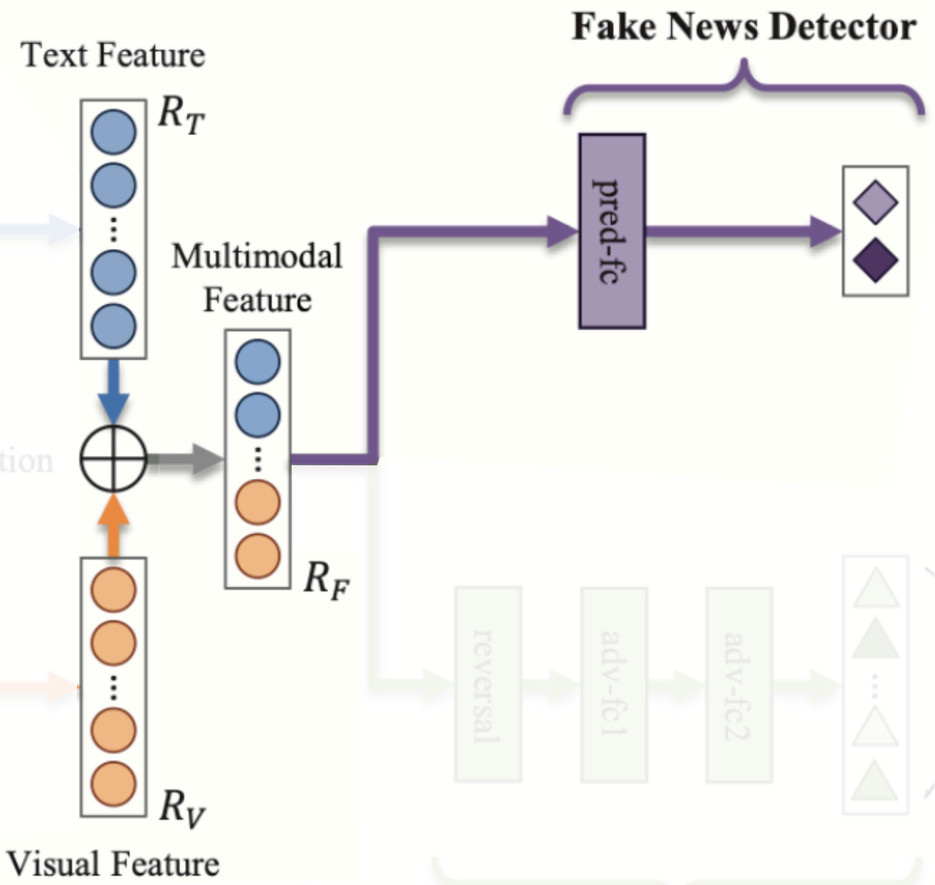
Fake News Detector



- Denote as $G_d(\cdot;\theta_d)$, θ_d: detector parameters
- Deploy a fully connected layer with softmax to predict the post are fake or real.
- **Probability** of post m_i being a fake one:
  - $P_\theta(m_i) = G_d(G_f(m_i;\theta_f);\theta_d)$
- Employ <u>**cross entropy**</u> to calculate the **detection loss**:
  - $L_d(\theta_f,\theta_d) = -\mathbb{E}_{(m,y)\sim(M,Y_d)}[y\log(P_\theta(m)) + (1-y)(\log(1-P_\theta(m))]$
- <u>Minimize loss</u> function by seeking the optimal parameters θ_f, θ_d
  - $(\hat{\theta}_f,\hat{\theta}_d) = arg\ \min_{\theta_f,\theta_d} L_d(\theta_f,\theta_d)$

Methodology.....

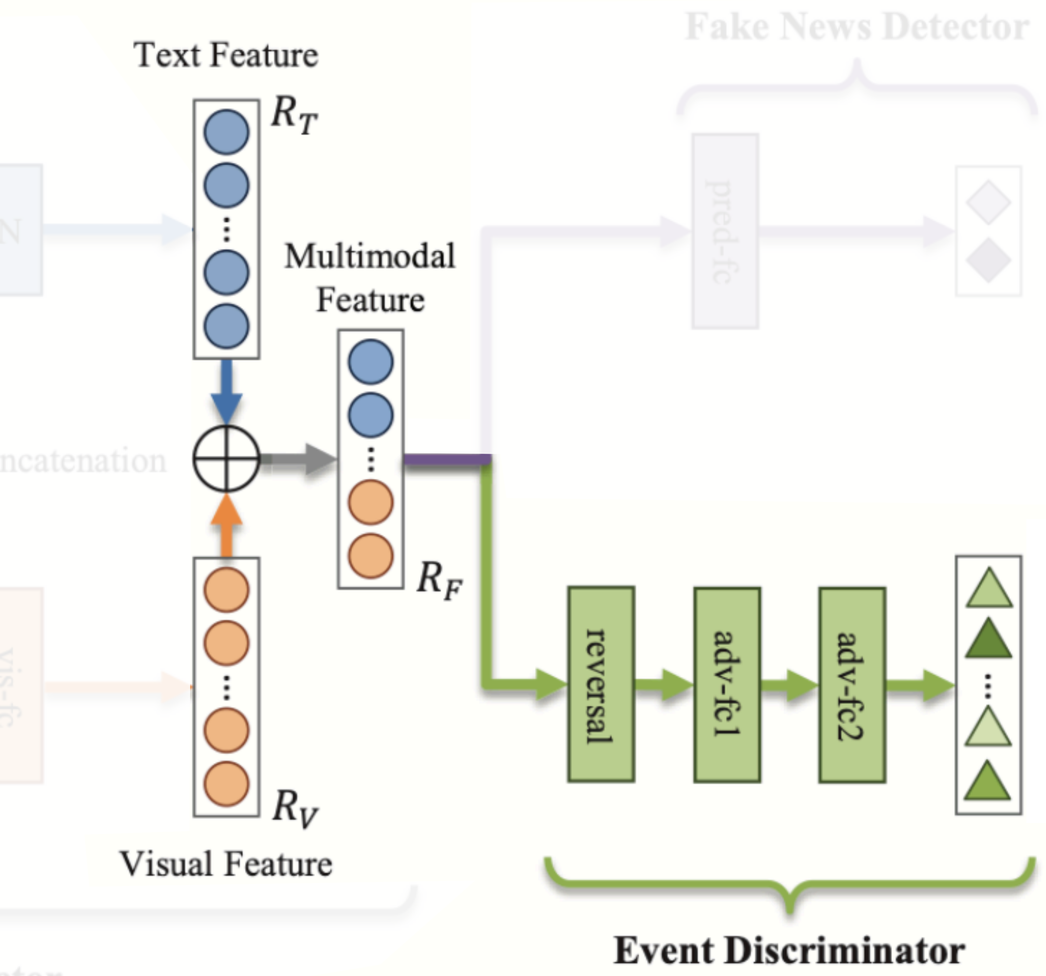
Fake News Detector



- As main goal of this work is detect the event which not covered by the training dataset.
- Direct minimization of detection loss only helps detect fake news included in the training dataset
  - Capture only event-specific knowledge or patterns
  - Not generalize well
- Need to learn more general feature representations that can capture the common features among all the events.
  - Should be event-invariant and doesn't include any event-specific feature.

Methodology.....

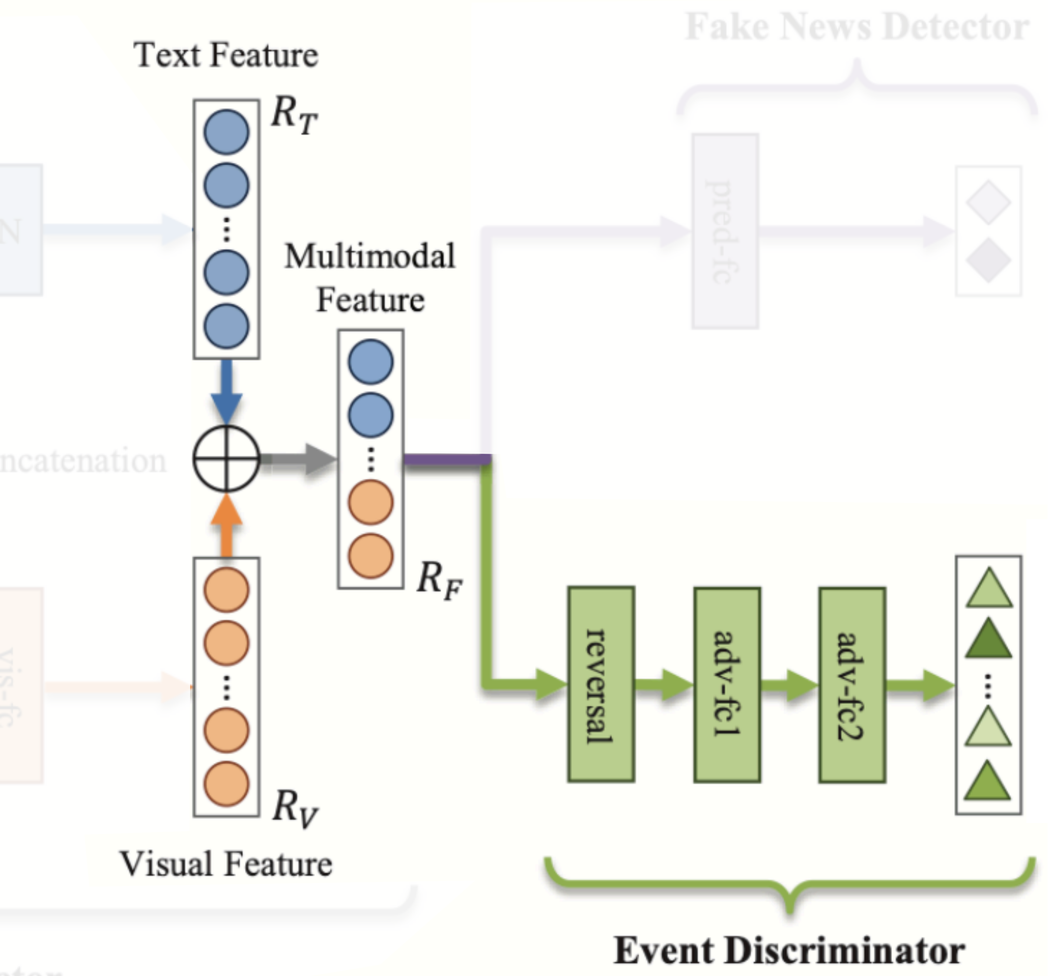
Event Discriminator



- Measure the dissimilarities of the feature representations among events and remove them in order to capture the event invariant feature representation.
- NN which consists of two fully connected layers with corresponding activation functions.
- Aims to correctly classify the post into one of K events based on the multi-modal feature representations.
- Denote as $G_e(R_F; \theta_e)$, θ_e: parameters

Methodology.....

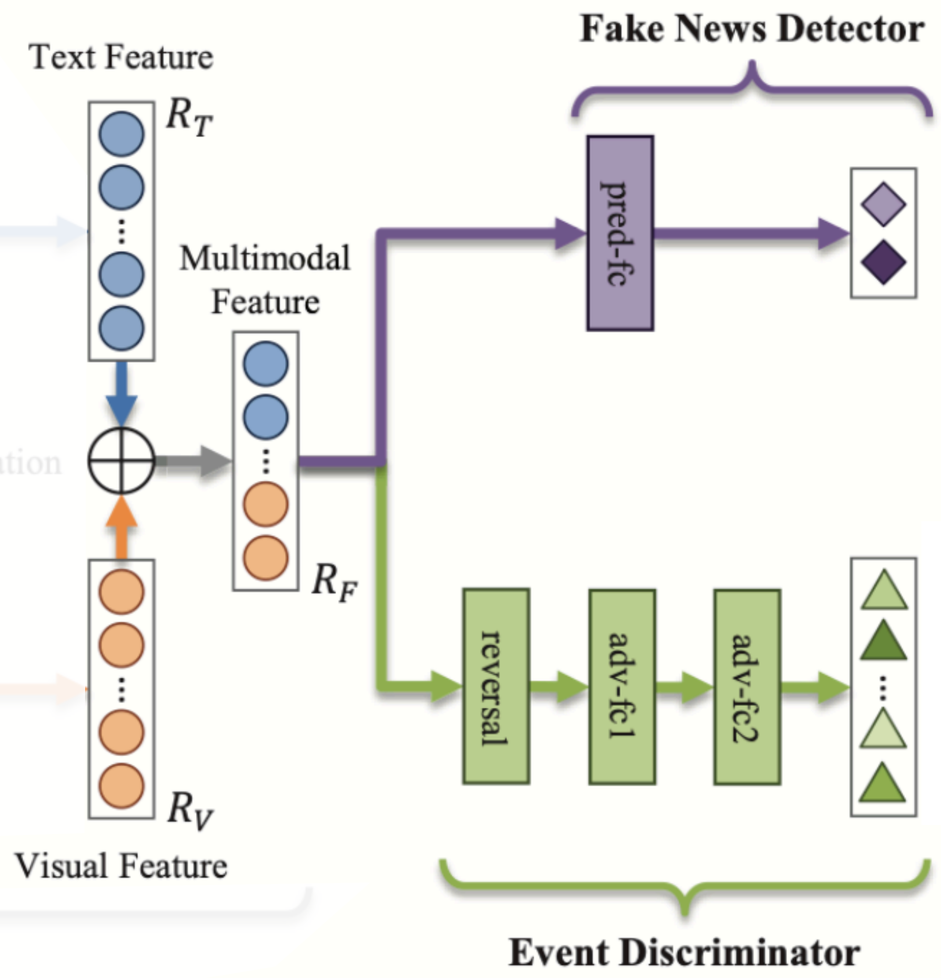
Event Discriminator



- Loss function by cross entropy:
  - $L_e(\theta_f, \theta_e) = -\mathbb{E}_{(m,y)\sim(M,Y_e)}[\sum_{k=1}^{K} 1_{[k=y]} \log(G_e(G_f(m;\theta_f);\theta_e))]$
- Parameters minimizing the loss $L_e(\cdot,\cdot)$:
  - $\hat{\theta}_e = arg \min_{\theta_e} L_e(\theta_f, \theta_e)$
- Large loss means the events' representations are similar and the learned feature are event-invariant.
  - Need to **maximize** the $L_e(\theta_f, \hat{\theta}_e)$ by seeking the optimal parameters θ_f

Methodology.....

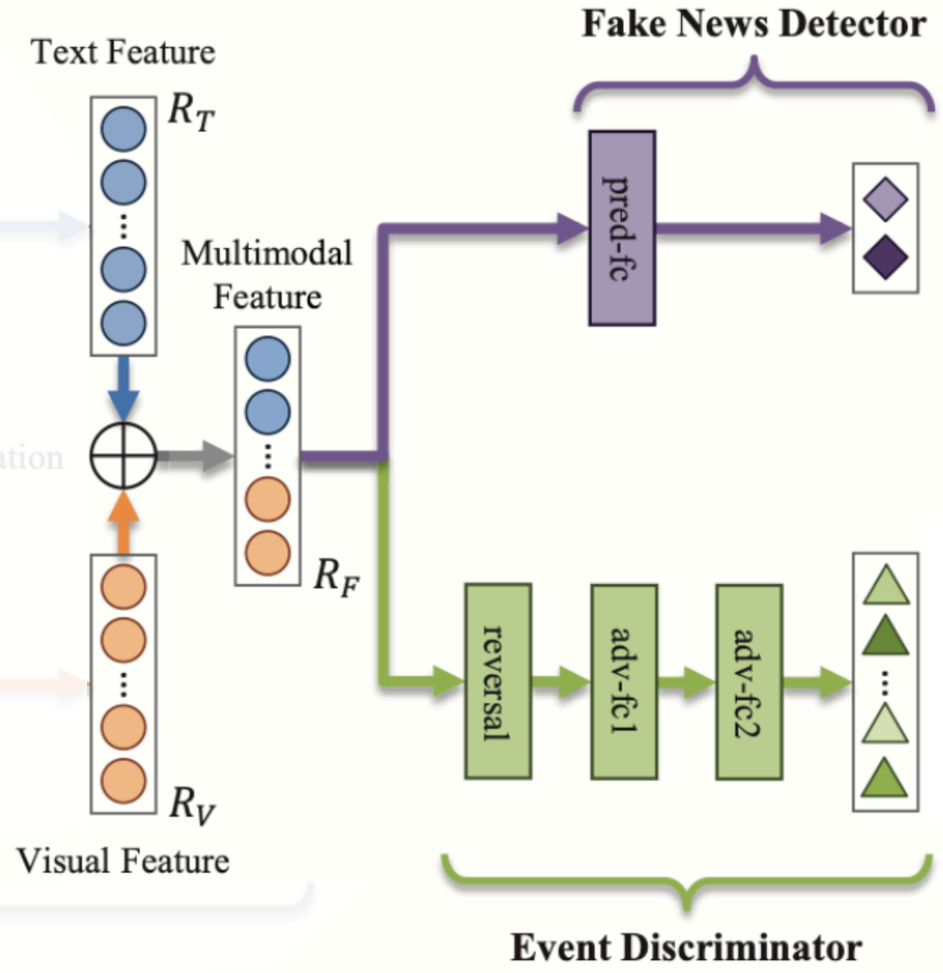
Model Integration



- $G_f(\cdot;\theta_f)$ need to cooperate with $G_d(\cdot;\theta_d)$ to **minimize the $L_d(\theta_f,\theta_d)$** to improve performance
- $G_f(\cdot;\theta_f)$ tries to fool $G_e(\cdot;\hat{\theta}_e)$ to achieve event-invariant representations by maximizing $L_e(\theta_f,\theta_e)$
- Define loss of this three-player game as
  - $L_{final}(\theta_f,\theta_d,\theta_e) = L_d(\theta_f,\theta_d) - \lambda L_e(\theta_f,\theta_e)$
  - In this paper, simply set $\lambda = 1$ to without tuning the trade-off parameter.

Methodology.....

Model Integration



- parameter set we seek is the saddle point of the final objective function, use SGD to solve problem:
  - $(\hat{\theta}_f, \hat{\theta}_d) = arg \min_{\theta_f, \theta_d} L_{final}(\theta_f, \theta_d, \hat{\theta}_e)$
  - $\hat{\theta}_e = arg \max_{\theta_e} L_{final}(\hat{\theta}_f, \hat{\theta}_d, \theta_e)$
- Here adopt the gradient reversal layer (GRL)
  - Acts as an identity function during forward stage, and it multiplies gradient with $-\lambda$ and passes the results to the preceding layer during back-prop stage.
  - GRL easily added between $G_f(\cdot\,; \theta_f)$ and $G_e(\cdot\,; \hat{\theta}_e)$

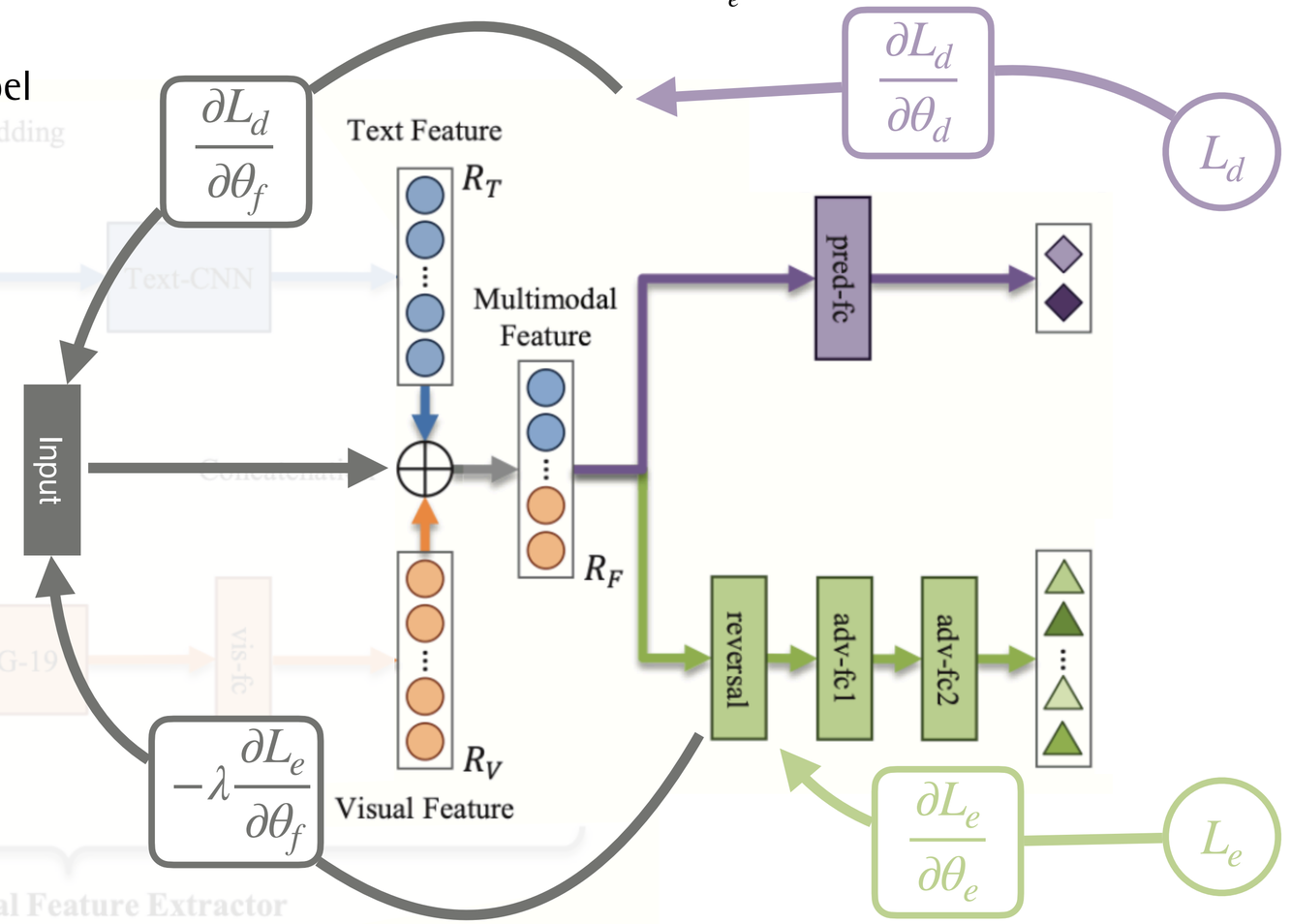
Methodology.....

Gradient Reversal Layer

- $L_{final}(\theta_f, \theta_d, \theta_e) = L_d(\theta_f, \theta_d) - \lambda L_e(\theta_f, \theta_e)$
- $(\hat{\theta}_f, \hat{\theta}_d) = arg \min_{\theta_f, \theta_d} L_{final}(\theta_f, \theta_d, \hat{\theta}_e)$
- $\hat{\theta}_e = arg \max_{\theta_e} L_{final}(\hat{\theta}_f, \hat{\theta}_d, \theta_e)$

Input: The multi-modal input $\{m_i\}_{i=1}^N$, the auxiliary event label $\{e_i\}_{i=1}^N$, the detection label $\{y_i\}_{i=1}^N$ and the learning rate η

1. **for** number of training iterations **do**
2. Decay learning rate: $\eta' = \frac{\eta}{(1 + \alpha \cdot p)^\beta}$
3. Update $\theta_f \leftarrow \theta_f - \eta(\frac{\partial L_d}{\partial \theta_f} - \lambda \frac{\partial L_e}{\partial \theta_f})$
4. Update $\theta_e \leftarrow \theta_e - \eta \frac{\partial L_e}{\partial \theta_e}$
5. Update $\theta_d \leftarrow \theta_d - \eta \frac{\partial L_d}{\partial \theta_d}$
6. **end for**



Experiments.

Dataset

Method	Twitter	Weibo
# of fake News	7898	4749
# of real News	6026	4779
# of image	514	9528

- **Twitter dataset**
- from MediaEval Verifying Multimedia Use benchmark
- Contain text, attach image/video and additional social context information
  - Focus on text and image
  - Remove the tweets without any text or image
- Has two parts: the **development and test set, there is no overlapping events among them.**

- **Weibo dataset**
- Fake news posts: 2012.05 ~ 2016.01 verified by Weibo official rumor debunking system
- Real news posts: 2012.05 ~ 2016.01 from Weibo verified by Xinhua News Agency
  - removed duplicated and very small images
- Use single-pass clustering and split whole dataset into **training, validation, testing sets = 7:1:2 to ensure that they don't not contain any common event.**

Experiments..

Baselines

- To validate the effectiveness of EANN, choose baselines from the following three cat.:
  - Single Modality Models
    - **Text / Vis**
  - Multi-Modal Models
    - **VQA / NeuralTalk / att-RNN**
  - Variant of the proposed Model
    - **EANN-** (w/o the event discriminator)

Experiments...

Performance Comparison: Twitter Dataset

Method	Accuracy	Precision	Recall	F1
Text	0.532	0.598	0.541	0.568
Vis	0.596	0.695	0.518	0.593
VQA	0.631	0.765	0.509	0.611
NeuralTalk	0.610	0.728	0.504	0.595
att-RNN	0.664	0.749	0.615	0.676
EANN-	0.648	0.810	0.498	0.617
EANN	**0.715**	**0.822**	**0.638**	**0.719**

- # of Tweets on different events is imbalanced and more than 70% of tweets are related to a single event.
  - Cause the learned the text feature mainly focus on some specific events.
  - Seriously prevent extracting transferable feature among events on **Text** Model
- **Text** is lowest, **Vis** is better than **Text**.
  - Images are more transferable, with VGG19 extracting useful feature.
  - **Vis** still worse than that multi-modal approaches
    - Confirms that multiple modalities is superior for the task of fake news detection.

Experiments....

Performance Comparison: Twitter Dataset

Method	Accuracy	Precision	Recall	F1
Text	0.532	0.598	0.541	0.568
Vis	0.596	0.695	0.518	0.593
VQA	0.631	0.765	0.509	0.611
NeuralTalk	0.610	0.728	0.504	0.595
att-RNN	0.664	0.749	0.615	0.676
EANN-	0.648	0.810	0.498	0.617
EANN	**0.715**	**0.822**	**0.638**	**0.719**

- **att-RNN** performs better than **VQA** and **NeuralTalk**
  - Shows that applying attention mechanism can improve
- **EANN-** tend to capture the event-specific features
  - Would lead failure of learning enough shared features among events
- **EANN** significantly improves the performance in terms of all the measures

Experiments.....

Performance Comparison: Weibo Dataset

Method	Accuracy	Precision	Recall	F1
Text	0.763	0.827	0.683	0.748
Vis	0.615	0.615	0.677	0.645
VQA	0.773	0.780	0.782	0.781
NeuralTalk	0.717	0.683	**0.843**	0.754
att-RNN	0.779	0.778	0.799	0.789
EANN-	0.795	0.806	0.795	0.800
EANN	**0.827**	**0.847**	0.812	**0.829**

- Similar result can be observed as those on Twitter dataset.
- However, we can see that **Text** is greatly higher than that of **Vis**
  - Because Weibo dataset doesn't have imbalanced issue, and with sufficient diversity, useful linguistic patterns can be extracted.
  - Images of Weibo dataset are more complicated in semantic meaning than Twitter.
  - **Vis** can't learn meaningful representations, although use VGG19 (But MVNN can?)

Experiments.....

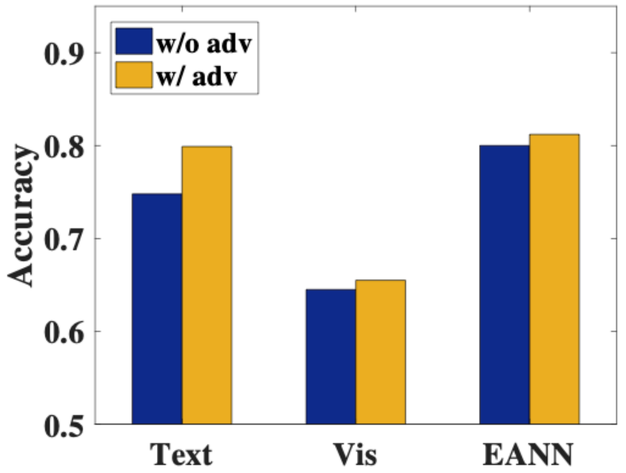
Performance Comparison: Weibo Dataset

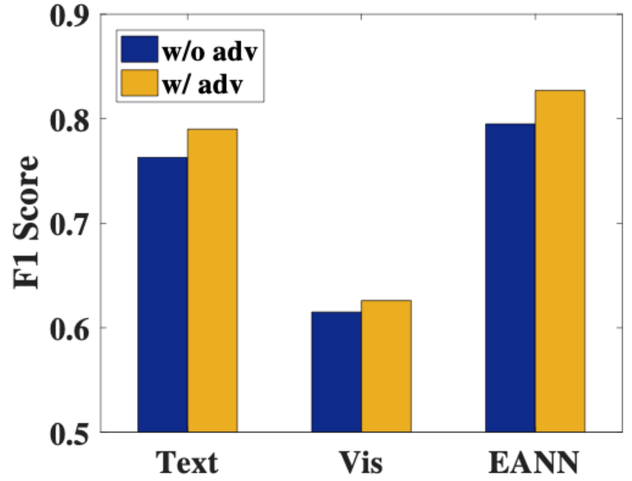
Method	Accuracy	Precision	Recall	F1
Text	0.763	0.827	0.683	0.748
Vis	0.615	0.615	0.677	0.645
VQA	0.773	0.780	0.782	0.781
NeuralTalk	0.717	0.683	**0.843**	0.754
att-RNN	0.779	0.778	0.799	0.789
EANN-	0.795	0.806	0.795	0.800
EANN	**0.827**	**0.847**	0.812	**0.829**

- **EANN-** is better than all multi-modal approaches on Weibo dataset
  - Since length of each post is relatively short (<140 characters), Text-CNN may capture more local representative features.
- **EANN** compared with **EANN-**
  - Can conclude that using event discriminator component indeed improves the performance of fake news detection

Experiments.....

Event Discriminator Analysis



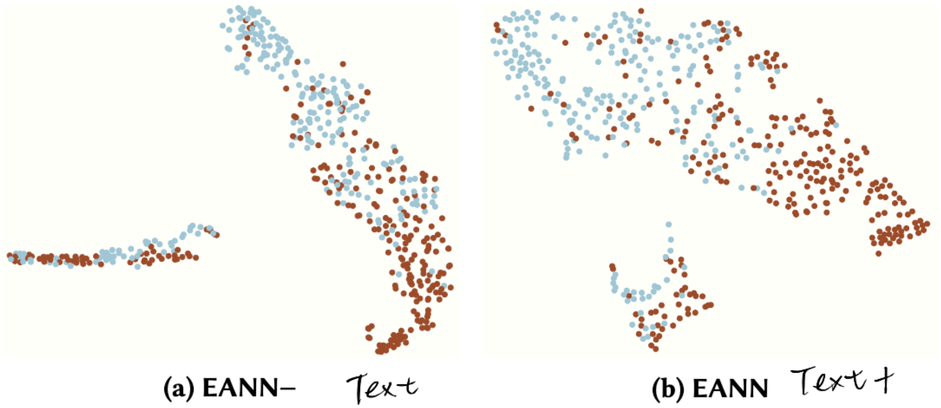


- Both accuracy and F1 score of **Text+** and **Vis+** are greater than those of **Text** and **Vis.**
- **EANN-** vs. **EANN** already discuss before
- Thus we can draw a conclusion that incorporating event discriminator is essential and effective for the task of fake news detection.

Only Text Feature (Only R_T)

Experiments.....

Event Discriminator Analysis



(a) EANN– Text (b) EANN Text+

- Qualitatively visualize the text features R_T learned by **EANN-** and **EANN** on Weibo testing set with t-SNE
- **EANN-** can learn discriminable features, but still twisted together, especially left of (a)
- **EANN** are more discriminable, and there are bigger segregate areas among samples with different labels.
  - Prove event discriminator is **effective** and thus **achieves better performance**

Experiments.....

Case Studies for Multiple Modalities: Text Missed

(a) Five headed snake

(b) Photo: Lenticular clouds over Mount Fuji, Japan. #amazing #earth #clouds #mountains

- The text content don't show evidence to identify that the tweets are fake.
- For both of the examples in Figure, they describe the image with common patterns.
- Text model identifies this news as a real one.
- As seen, two attached images look quite suspicious and are very likely to be forged pictures.
- By feeding visual and texture into EANN, both tweets are classified as fake with high confidence scores.

Experiments.....

Case Studies for Multiple Modalities: Image Missed

(a) Want to help these unfortunates? New, Iphones, laptops, jewelry and designer clothing could aid them through this!

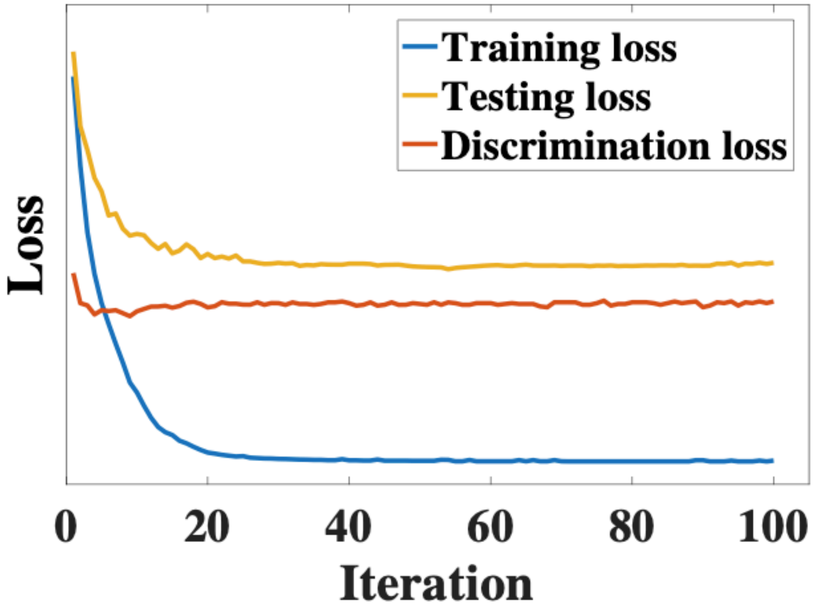
(b) Meet The Woman Who Has Given Birth To 14 Children From 14 Different Fathers!

- For first example, the complicated semantic meaning is contained in the attached image, which is challenging to be captured by VGG19.
- However, the words with strong emotion and inflammatory intention suggest this is a suspicious post.
- Second example looks very normal, but the text content seems to misrepresent the image to mislead the readers.
- Without texture content, the meaning of tweets would totally change.
- Only aligned with the corresponding text description, it can be identified as fake news.

Experiments.....

Convergence Analysis

- At the beginning, all of losses decrease.
- Then the discrimination loss increases and stabilizes at a certain level.
  - Decreasing in the beginning represents the event discriminator detecting the event-specific info in the feature.
  - the feature representation tend to be event invariant by the minimax game, specific info is removed incrementally, and the discrimination loss increases over the time.
- Then all losses smoothly converge, means that a certain level of equilibrium have been achieved.



Conclusions and Contribution

- Study the problem of multi-modal content fake news detection
- Overcome the major challenge of fake news detection stems from newly emerged events on which existing approaches only showed unsatisfactory performance.
- First to propose a novel Event Adversarial Neural Network framework which can learn transferable features for unseen events.
- EANN models is a general framework, can be easily replace by different model designed for feature extractions.

Comments

- Focus on adversarial neural network
  - Learned event-invariant representation via event discriminator
- Can apply with other work via replace the feature extractor
- Only use concatenation fusing texture and visual content features
  - Maybe can try use attention mechanism to fuse for stronger feature representation