Exploiting Multi-domain Visual Information for Fake News Detection

P. Qi, J. Cao, T. Yang, J. Guo and J. Li,

2019 IEEE International Conference on Data Mining (ICDM)

20210304 何嘉峻

Fake-news image.

Fake-news image cat.

- Classify fake-news images into two cat.:
  - tampered images
  - misleading images

(a) Tampered Images

(b) Misleading Images

Fake-news image..

Physical level images

- Re-compressed and tampered images often present periodicity in the **frequency domain**
  - design a **CNN-based network** to automatically capture the characteristics of images in the frequency domain.

(a) A fake-news image

(b) A real-news image

Fake-news image...

Semantic level images

- Exhibit some distinct characteristics in the **pixel domain**, visual factors from low-level to high-level
  - build a **multi-branch CNN-RNN network** to extract features of different semantic levels for fully capturing the characteristics of fake-news images in the pixel domain.

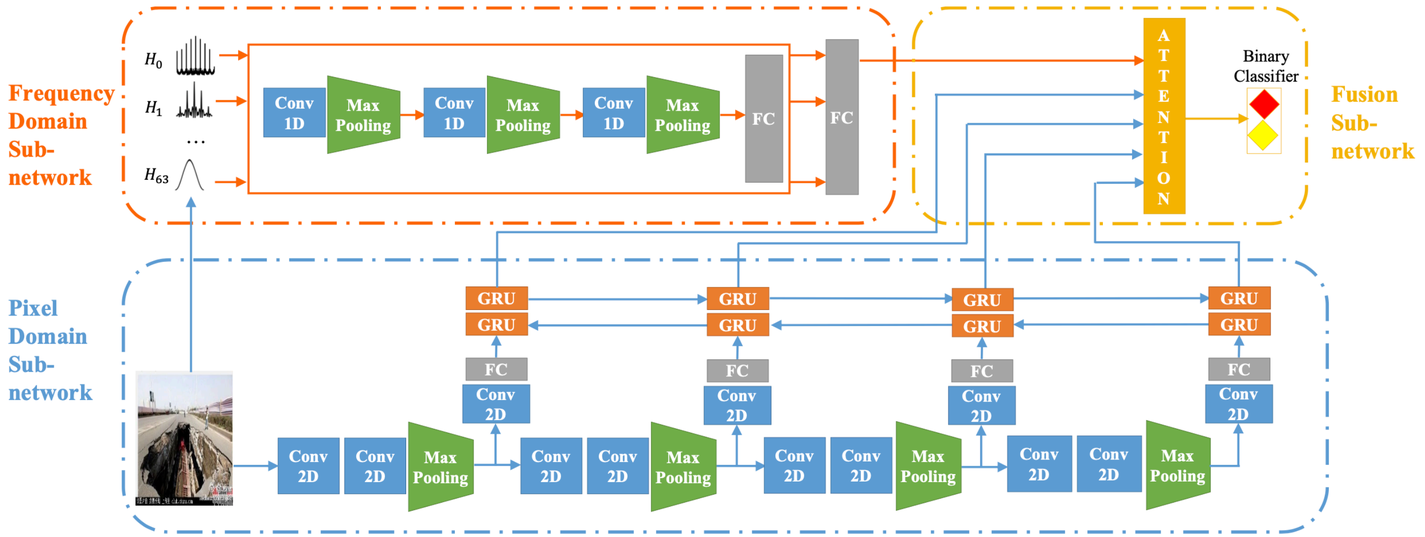
Attention Mechanism

- Visual features at both physical and semantic levels are important for detecting fake-news images
- Not all features contribute **equally** to the task of fake news detection
  - some visual features play more important roles than others in evaluating feature
- Employ an **attention mechanism to fuse** these visual features from frequency and pixel domains dynamically.

Introduction

MVNN framework

- To sum up, propose a Multi-domain Visual Neural Network (MVNN) framework, which can learn effective visual representations by combining the information of frequency and pixel domains for fake news detection.
- model consists 3 main components:
  - a frequency domain sub-network
  - a pixel domain sub-network
  - a fusion subnetwork



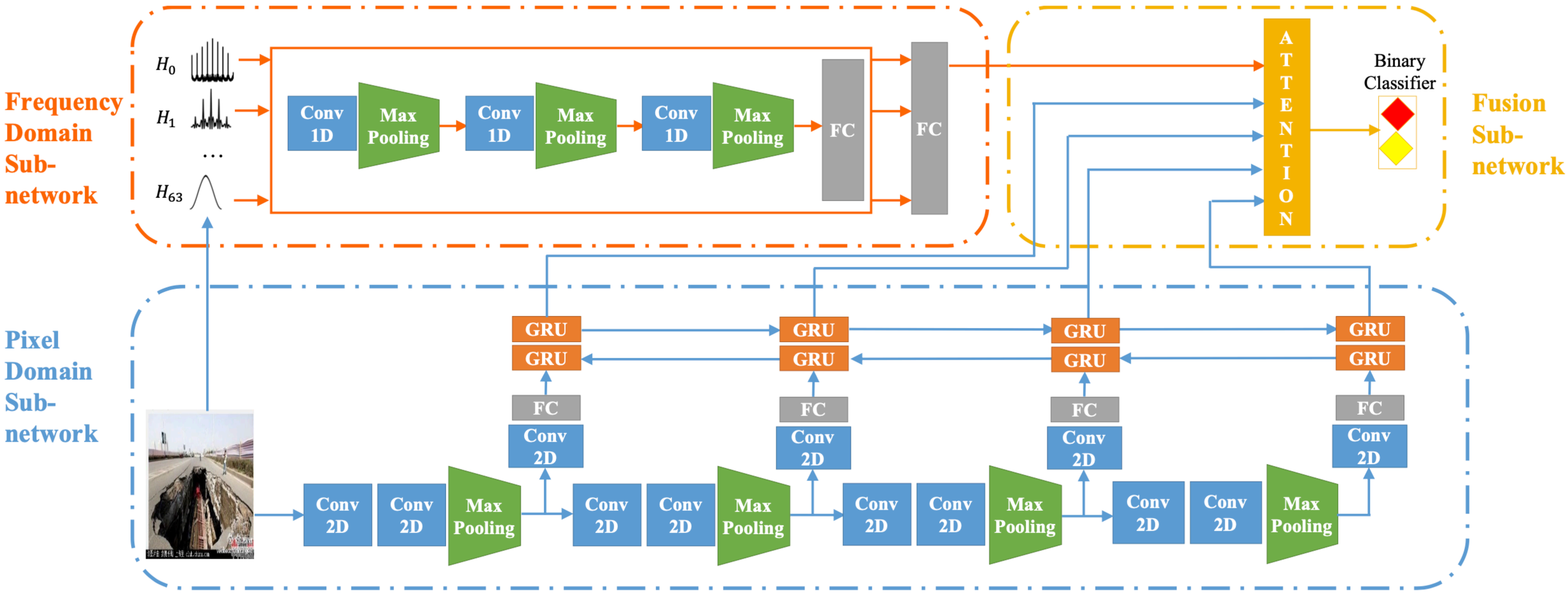
Problem Formulation

Fake-news Definition

- *Definition 1:* **Fake news:** In the context of microblog, a piece of fake news is a news post that is intentionally and verifiably false.
- *Definition 2:* **Fake-news images:** A fake-news image is an image attached to fake news.
- **Problem 1:** *Given a set of news posts* $X = \{x_1, x_2, \ldots, x_m\}$, *corresponding images* $I = \{i_1, i_2, \ldots, i_m\}$, *and labels* $Y = \{y_1, y_2, \ldots, y_m\}$, *learn a classifier* f *that can utilize the corresponding image to classify whether a given post is fake news* $(y_t = 1)$ *or real news* $(y_t = 0)$, *i.e.,* $\hat{y}_t = f(i_t)$.

Methodology.

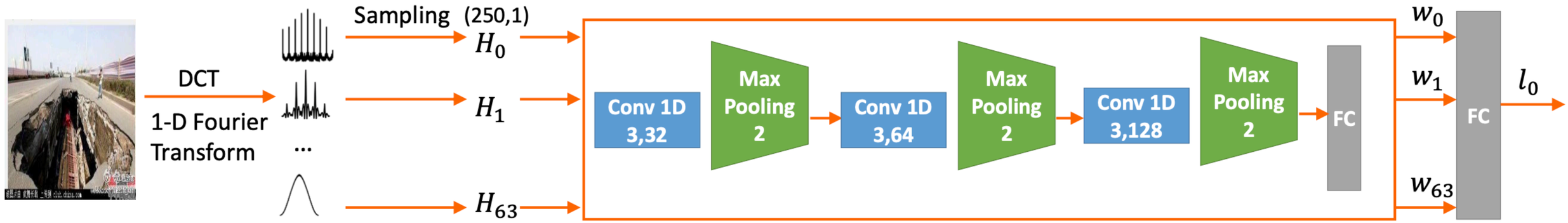
Model Overview



Methodology..

Frequency Domain Sub-network

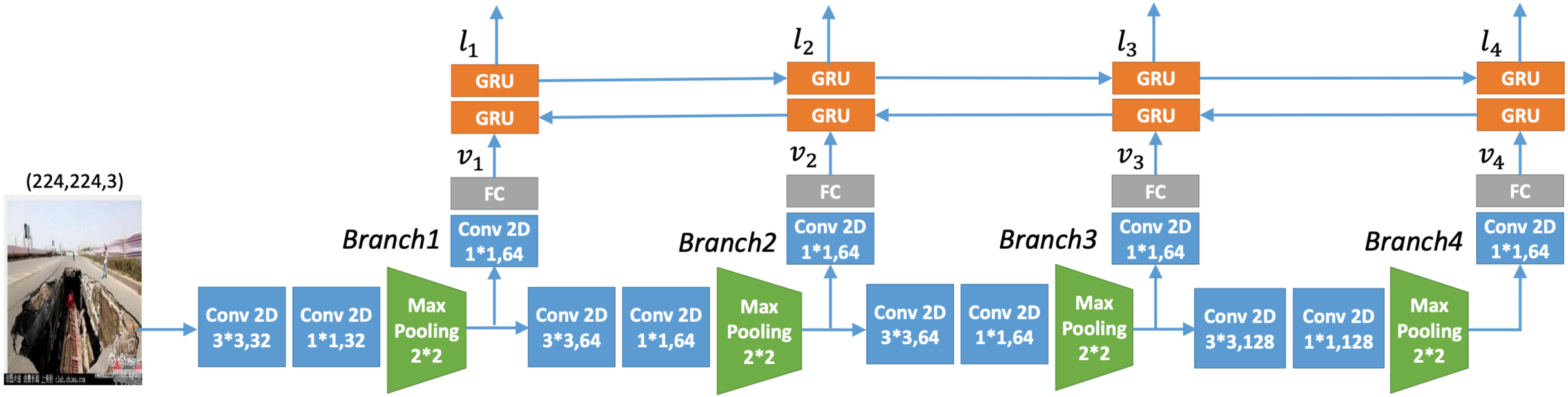
- Frequency domain sub-network to extract features of physical level images → l_0



Methodology...

Pixel Domain Sub-network

- Pixel domain sub-network to extract features of semantic level images → $\{l_t\}, t \in [1,4]$



Methodology....

Fusion Sub-network

- Frequency domain sub-network (physical level): l_0
- Pixel domain sub-network (semantic level): $\{l_1, l_2, l_3, l_4\}$
- via attention mechanism, and the enhanced image representation is computed as follows:

  - $F(l_i) = v^T \tanh(W_f l_i + b_f), i \in [0,4]$, $\alpha_i = \frac{\exp(F(l_i))}{\sum_i \exp(F(l_i))}$, $u = \sum_i \alpha_i l_i$

Methodology.....

Fusion Sub-network

- $F(l_i) = v^T \tanh(W_f l_i + b_f), i \in [0,4]$, $\alpha_i = \frac{\exp(F(l_i))}{\sum_i \exp(F(l_i))}$, $u = \sum_i \alpha_i l_i$
  - W_f: weight matrix, b_f: bias term, v^T: transposed weight vector
  - F: score function, a_i: normalized weight of i-th feature vector
- Obtained high-level representation of image u at both physical and semantic levels.
- Use a fully-connected layer with softmax activation to project to two classes, gain the **probability distribution:** $p = \text{softmax}(W_c u + b_c)$

Methodology.....
Fusion Sub-network

- Loss function
  - $L = -\sum [y \log p + (1 - y)\ log(1 - p)]$
  - y: ground truth: 1= fake-news images, 0 = real news images
  - p: predicted probability of being fake-news images

Experiments.

Dataset

- **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
  - text-only posts are removed and only one image is saved for posts
- In total, this dataset includes 4749 fake news posts and 4779 real news posts
- First use K-means algorithm to cluster all news into 200 clusters, and split whole dataset into **training set : validation set : testing set = 7:1:2**

Experiments...

Baselines

- To validate the effectiveness of MVNN, choose several representative methods which are used to model the visual contents for fake news detection as baselines:
  - **Forensics Features (FF)+LR (MediaEval, 2015)**
  - **Pre-trained VGG**
  - **Fine-tuned VGG**
  - **ConvAE (ICANN, 2011)**

Experiments....

Evaluation questions

- To evaluate the effectiveness, aim to answer the following evaluation questions:
  - **EQ1:** Is MVNN able to improve the performance of fake news detection based on visual modality?
  - **EQ2:** How effective are different domains and other network components: attention, Bi-GRU and branches in the pixel domain sub-network, in improving the performance of MVNN?
  - **EQ3:** Can MVNN help improve the performance of multi-modal fake news detection?

Experiments.....

Ablation Study

Method	Accuracy	Precision	Recall	F1
FF+LR	0.650	0.612	0.579	0.595
Pre-trained VGG	0.721	0.669	0.738	0.702
Fine-tuned VGG	0.754	0.74	0.689	0.714
ConvAE	0.734	0.685	0.744	0.713
MVNN	**0.846**	**0.809**	**0.857**	**0.832**

- **EQ1:** Is MVNN able to improve the performance of fake news detection based on visual modality?

- 1) MVNN is best, validates MVNN can effectively capture the intrinsic characteristics of fake-news images, achieves an accuracy of 84.6%, outperforming existing approaches by 9.2%

- 2) Fine-tuned VGG better than Pre-trained VGG, show that the learned features are more relevant to the task of fake news detection after fine-tuning the model on the fake news dataset.

Experiments.....

Ablation Study

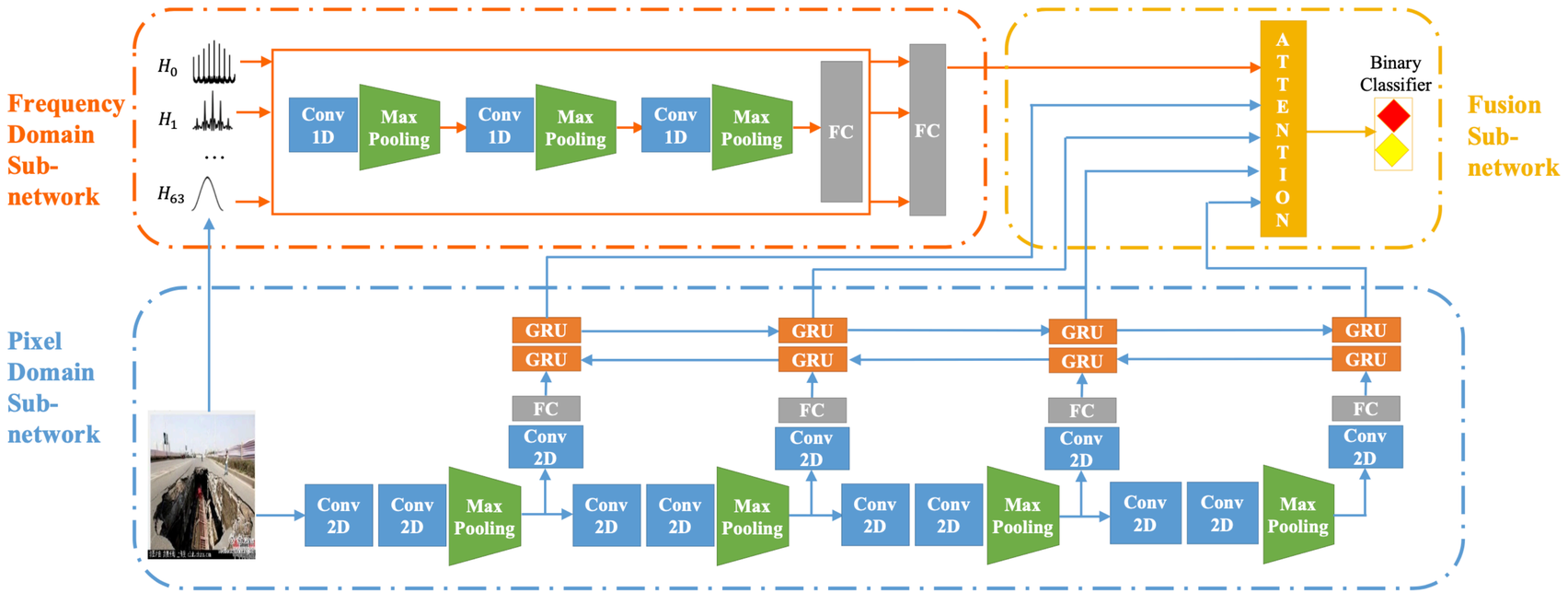
Method	Accuracy	Precision	Recall	F1
FF+LR	0.650	0.612	0.579	0.595
Pre-trained VGG	0.721	0.669	0.738	0.702
Fine-tuned VGG	0.754	0.74	0.689	0.714
ConvAE	0.734	0.685	0.744	0.713
MVNN	**0.846**	**0.809**	**0.857**	**0.832**

- **EQ1:** Is MVNN able to improve the performance of fake news detection based on visual modality?
- 3) Performance of ConvAE is slightly better than Pre-trained VGG. Show that ConvAE has the ability of understanding universal semantics of images, similar to models pre-trained in a supervised manner.
- 4) Performance of FF+LR is the worst methods because captured forensics features is very limited

Experiments.....

Ablation Study

- **EQ2:** How effective are different domains and other network components: attention, Bi-GRU and branches in the pixel domain sub-network, in improving the performance of MVNN?
- To illustrate the effectiveness of different domains and other network components by removing certain components:
  - w/o frequency domain
  - w/o pixel domain
  - w/o attention
  - w/o Bi-GRU
  - w/o branches



Experiments.....

Ablation Study Observations

Method	Accuracy	Precision	Recall	F1
MVNN	**0.846**	**0.809**	**0.857**	**0.832**
w/o frequency domain	0.794	0.792	0.728	0.758
w/o pixel domain	0.737	0.698	0.717	0.708
w/o attention	0.827	0.778	0.853	0.814
w/o Bi-GRU	0.828	0.772	0.841	0.805
w/o branches	0.803	0.752	0.830	0.789

- **Multiple domains:** The frequency and pixel domain both are important, the accuracy drops by 5.2% and 10.9% without the frequency and pixel domain sub-network. Pixel domain plays a major role and the frequency domain is auxiliary.

- **Network Components:**

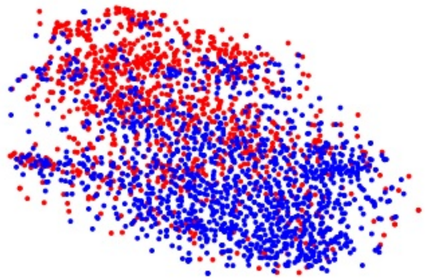
  - remove attention the accuracy is drops by 1.9%, which means that the attention mechanism better than simply concatenating

  - remove the Bi-GRU reduces 1.8%; remove the branches drops by 4.3%.

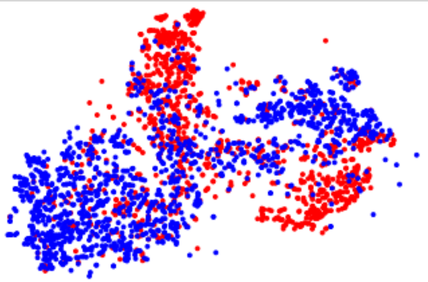
- Incorporating different levels of features and considering the dependencies between these features both help capture the semantic characteristics of visual contents

Experiments.....

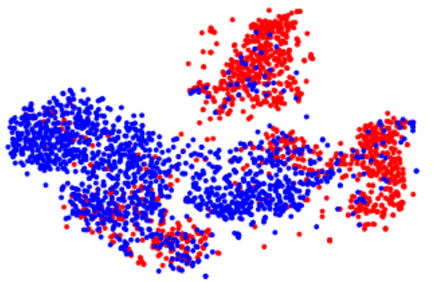
visualize the visual features



(a) Frequency domain sub-network



(b) Pixel domain sub-network



(c) MVNN

- t-SNE show separability of the feature representations: MVNN > pixel > frequency
  - **frequency domain:** positive and negative feature samples overlap a lot
  - **pixel domain:** can learn discriminable features, but the learned features are still twisted together
  - **MVNN:** there is a relatively visible boundary between samples with different labels
- Pixel domain is more effective than frequency domain in distinguishing
- **Fuses information of multiple domains can more distinctive** feature representations, better than single domain

Experiments.....

Application on Multi-modal Fake News Detection

- **EQ3:** Can MVNN help improve the performance of multi-modal fake news detection?
- experiment with three fusing methods as follows:
  - **attRNN (ACM MM, 2017)**
  - **EANN (ACM SIGKDD, 2018)**
  - **MVAE (ACM WWW, 2019)**

Experiments.....

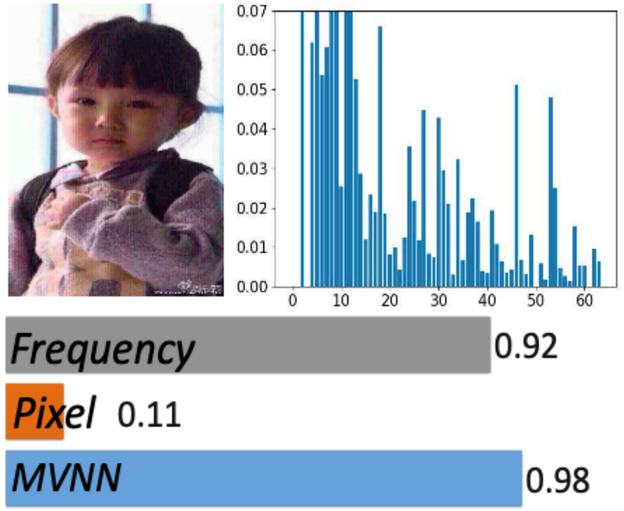
Application on Multi-modal Fake News Detection

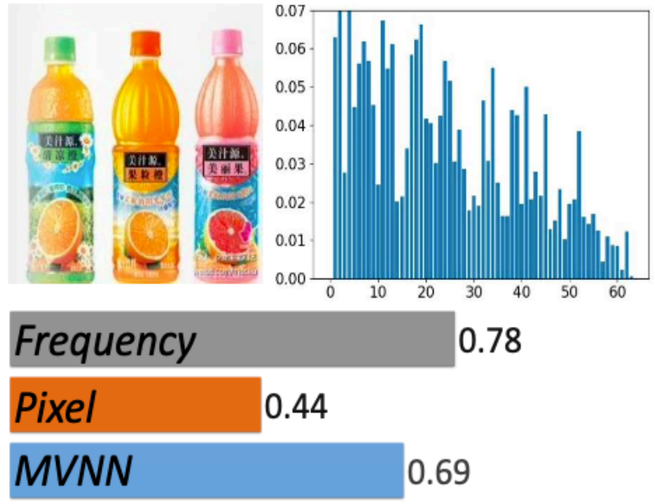
- MVNN consistently outperforms other baselines
  - MVNN by over 5.2% in accuracy, MVNN can easily replace existing methods to obtain the representations of visual contents
- FF+LR in attRNN is obviously worse than EANN and MVAE
- attRNN can hardly utilize the semantic alignment between the text and the forensics features to fuse the textual and visual information

	Method	Accuracy	Precision	Recall	F1
attRNN	FF+LR	0.735	0.801	0.665	0.727
	Pre-trained VGG	0.821	0.813	0.862	0.837
	Fine-tuned VGG	0.849	0.888	0.818	0.852
	ConvAE	0.816	0.848	0.796	0.821
	MVNN	**0.901**	**0.911**	**0.901**	**0.906**
EANN	FF+LR	0.780	0.840	0.724	0.778
	Pre-trained VGG	0.821	0.861	0.791	0.824
	Fine-tuned VGG	0.841	0.883	0.807	0.843
	ConvAE	0.823	0.863	0.794	0.827
	MVNN	**0.897**	**0.930**	**0.872**	**0.900**
MVAE	FF+LR	0.777	0.776	0.815	0.795
	Pre-trained VGG	0.813	0.893	0.737	0.804
	Fine-tuned VGG	0.832	0.875	0.798	0.835
	ConvAE	0.827	0.831	0.847	0.839
	MVNN	**0.891**	**0.896**	**0.898**	**0.897**

Case Studies

captured by MVNN
but missed by pixel domain

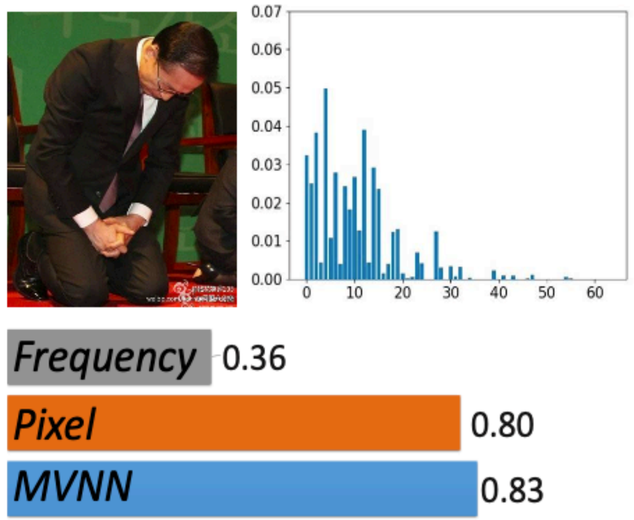


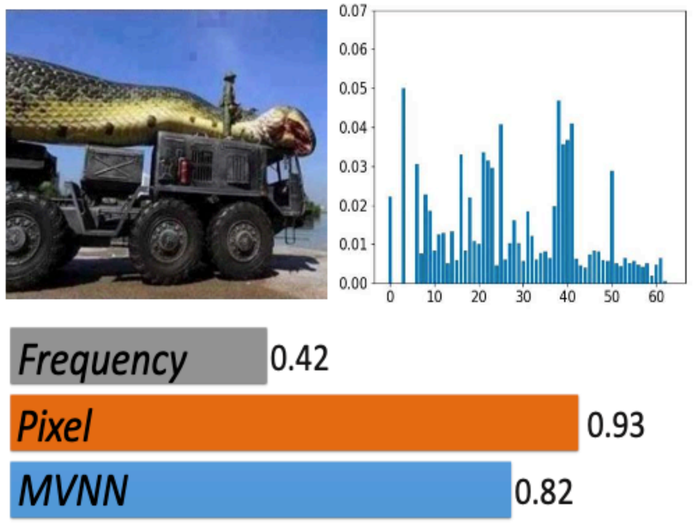


- **Frequency:** their frequency histograms look quite suspicious, showing that they are very likely to be **outdated images**.
- **Pixel:** the two images do not show evidence of fake news
- These two examples are correctly classified by MVNN while the results would totally change without the frequency information

Case Studies

captured by MVNN
but missed by frequency domain





- **Frequency:** histograms in Fig show little evidence of fake news
- **Pixel:** the image contents are eye-attracting and rather dubious
- By combining the information of the frequency and pixel domain, MVNN can easily detect that this is a fake-news image with high confidence.

Conclusions and Contribution

- Propose a framework MVNN to model the visual contents for fake news detection
  - exploits an end-to-end neural network to learn representations of frequency and pixel domains **simultaneously and effectively fuse them**
- Experiments conducted on Weibo dataset validate the effectiveness of MVNN, The results shows that MVNN is much better than existing methods.
- The visual representations learned by MVNN can help **improve the performance** of multi-modal fake news detection by a large margin.
- Proven the information of frequency and pixel domains are complementary

Comments

- Focus on the visual feature in this work
- Fuse the frequency and pixel feature with attention get stronger image representation
- Suitable for many types of fake-news image than other works
- Can apply with other work via replace the visual representation