

Preserving True Web Information using Multi-Modal Learning

Komal Krishna Mogilipalepu

December 13, 2024

Abstract

Detecting fake news and misleading information in social media, and preserving the true content in the web needs understanding of image and text together. Recent advancements in deep learning developed models to extract semantic information and compare image-text pairs. In this work, we are interested in adapting the developed large deep learning models to solve the problem of fake news detection. In particular, we review the work of [1], which tackles the problem of fake news detection by utilizing the resources of the web and the large image-text pre-trained model Contrastive Language-Image Pre-training (CLIP). The proposed method in [1] is able to identify the pristine and falsified image-caption pairs compared to the baselines. The hypothesis is that the proposed process in [1] can generalize to detect the image-caption pairs generated by deep learning models and also identify the modified true content by comparing it with the original information gathered from the web.

1 Introduction

Original images attached with fake text spread easily in the social media platforms and there is a high chance of believing that information [2]. It is important for social media platforms to verify the information before uploading them. To solve this problem, it is crucial to develop models, that can handle both image and text. Towards this, image-text matching and cross modal retrieval are two important problems to address the above challenge. Recently, deep learning models have shown best performance to extract the spatial information from images and temporal information from language. The extracted features are very useful in many supervised and unsupervised learning tasks. After learning the representations from image and text using the deep learning models, one should find the similarity between the two representations to know the relevance of image and text. Based on the similarity score the model decide whether the given image and text are a valid pair or not [3]. Once we have a model that identifies the valid image-text pair, one can use that model to detect the fake

news. In addition, image-text matching model, [1] use the resources from web to identify the fake news. For instance, gather evidence captions/images from the internet for the given query image/caption and evaluate the gathered information to identify the fake news. The method described in [1] to detect the fake news is explained in the following section.

2 Dataset

This paper consider the NewsCLIPPings dataset that consists of pristine and falsified image-caption pairs generated from the VisualNews dataset consists of image-captions from four news channels The Guardian, BBC, USA Today, and The Washington Post. For each pristine pair (img1, cap1) query for a threat pair (img2, cap2) based on the semantics, person, scene, SBERT-WK, and ResNet Place to generate out-of-context pair (img2, cap1). The generated falsified image-captions which are not close to the pristine one are filtered using adversarial CLIP image-text similarity score.

3 Methodology

Given the image and caption query, first gather the evidence from the internet, that is collect the caption evidence for the image query and collect the image evidence for the caption query. After that, find the visual reasoning between the query image and evidence images, similarly textual reasoning between the query caption and evidence captions, and visual-textual reasoning using the pre-trained CLIP model, and concatenate the three reasoning's to an embedding vector. The resultant embedding vector is normalized and passed through a two layer fully connected network with ReLu activation in the 1st layer and sigmoid in the 2nd layer that predicts the given image-caption query is pristine or falsified. The network is trained using a binary cross-entropy loss between the predicted and ground truth labels. [1] uses NewsClippings dataset, which contains 71,000 training, 7000 validation, and 7000 test image-caption pristine and falsified pairs.

3.0.1 Gathering Evidence

For each image-caption pair given in the NewsClippings dataset, the evidence is gathered from the Google vision API and Google custom search API. Given a query image \mathbf{I}_q the method collect the textual evidence consists of list of entities $\mathbf{E} = \{\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_M\}$ and list of captions/sentences $\mathbf{S} = \{\mathbf{S}_1, \mathbf{S}_2, \dots, \mathbf{S}_N\}$ using the Google custom search API, similarly, given a query caption \mathbf{C}_q the method collect the visual evidence $\mathbf{I}^e = \{\mathbf{I}_1^e, \mathbf{I}_2^e, \dots, \mathbf{I}_k^e\}$ using the Google vision API.

3.0.2 Checking Evidence

The gathered image evidence is compared with the query image by visual reasoning network, similarly the textual evidence is compared with the query caption using the textual reasoning network and also the query image and caption similarity is also evaluated with the Contrastive Language Image Pre-trained (CLIP) model. The three reasonings are concatenated to classify the given pair as pristine or falsefied. To get the visual reasoning between a query image and the evidence images, first get the features of the query image and evidence images using pre-trained ResNet model, then use those features to find the number of overlapped regions between the each evidence image and a query image. Similarly, to get the textual reasoning between a query caption and evidence captions, first get the embedding features of the each evidence sentence and a query caption by passing through a sentence embedding model and then use those embeddings to compare the query caption with the evidence sentences. To get the cross-modal reasoning pass the image and caption query through a CLIP model and find the similarity between image regions and words of a caption.

3.0.3 Classifier

The three reasoning embeddings such as visual, textual, and cross-modal are concatenated to get a resultant embedding. The resultant embedding is passed through a two fully connected layers with ReLU and sigmoid activations, and applied to a binary cross-entropy loss to get the output as pristine or falsefied.

4 Hypothetical Results

Figure 1 ,taken from [1], contains two examples of pristine and three examples of falsified from the NewsClippings dataset. In all the examples, to classify a given image-caption query as pristine/falsefied, the introduced model combines the textual and visual reasoning found from the gathered textual and visual evidences, and image-caption similarity using CLIP, which is fed to a classifier. As seen in the figure 1, in the first example, the model has given highest attention to the evidence that are most relevant to the query by following the names, entities, and semantic information to classify it correctly as pristine. In the second example, even though the background information for the image query and gathered images are different, the model still able to classify the image-caption correctly by giving more attention to the semantic meaning. In the third example, the given image-caption query appealing as a pristine example, however, the textual evidence and visual evidence is not having any relevance, therefore the model classified as correct label, which is falsified. In the fourth example, the textual evidence gives more attention to the contradicting cities and entities with same syntactic caption. Even though the image evidence is similar to the caption, still the model able to classify the query correctly as falsified. In the last example, the model incorrectly classify the pristine

example as falsified because the textual evidence has little connection with the query caption even though the context is same and the visual evidences are not similar to the query image.


















5 Conclusion

In this report, we discussed the importance of image-text matching in preserving the original information by detecting the unrealistic content. For example, we discussed the method and results of [1], which detects the fake news using the deep learning models. We hypothesize that the discussed method is a feasible approach to give highest priority to the original information exists in the web.

References

- [1] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz, “Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14940–14949.
- [2] Giandomenico Di Domenico, Jason Sit, Alessio Ishizaka, and Daniel Nunan, “Fake news, social media and marketing: A systematic review,” *Journal of Business Research*, vol. 124, pp. 329–341, 2021.
- [3] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang, “Multimodal fake news detection via clip-guided learning,” in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2825–2830.

Figure 1: Examples of fake news detection taken from [1]

Image-caption pair	Textual evidence	Visual evidence
 The Futenma marine corps airbase on the southern Japanese island of Okinawa	'United States', 'Ginowan', 'Governor', 'Military base', 'Politics', 'Japan', 'Takeshi Onaga', 'Governor of Okinawa Prefecture', 'Hiromasa Nakaima', 'Shinzo Abe', 'Okinawa', 'airport' 1- Hercules aircraft parked on the tarmac at Marine Corps Air Station Futenma in Ginowan on Okinawa; 2- Japan Decides to Stop Works on US Airbase Relocation in Okinawa; 3- Japan Decides to Restart Relocation of US Base in Okinawa Despite Protests.	  
 The soaring number of Syrian refugees has sparked increasing resentment in Lebanon	'Syria', 'Lebanon', 'United Kingdom', 'Tent', 'Syrians', 'Language', 'Refugee', 'Recreation', 'Tourism', 'Camping', 'Language barrier', 'rural area' 1- Syrian refugees at a camp in eastern Lebanon, December 2014; 2- Syrian entering Lebanon face new restrictions; 3- Among those displaced, 1.6 million children have fled Syria; 4- Syrian refugees in the UK: 'We will be good people. We will build this country'	  
 Healthcare activists say the ruling against Novartis ensures poor people will be able to access cheap versions of cancer medicines	'United States Capital', 'Affordable Care Act', 'Supreme Court of the United States', 'Presidency of Donald Trump', 'President of the United States', 'United States', 'us capitol grounds' 1- Demonstrators from Doctors for America in support of Obamacare march in front of the Supreme Court on March 4, 2015; 2- The Affordable Care Act Is Back In Court, 5 Facts You Need To Know; 3- As Court Hears Arguments in Lawsuit To Eliminate Obamacare, Conn. Senators Plead Their Case.	  
 Smoke rises following an Israeli air strike in Gaza City	'Kobane', 'Kurdistan Region', 'United States', 'Peshmerga', 'Turkey', 'Kurds', 'Syria', 'Iraq', 'kobani war' 1- Smoke rises after a U.S.-led airstrike in the Syrian town of Kobani 2- The border town of Kobani is under threat after the Islamists drove 180,000 Kurds into Turkey; 3- Former Kurdish Sniper Claims To Have Killed Around 250 ISIS Fighters.	  
 How can our young readers persuade their parents to get them a Playstation 3	'Grand Theft Auto V', 'Gamer', 'Grand Theft Auto IV', 'Wii', 'Grand Theft Auto V', 'PlayStation 3', 'Rockstar Leeds', 'very seaborne marshall', 'Gordon Hall', 'Rockstar Games' 1- A court order banning Sony from importing PS3s into the Netherlands has been lifted; 2- Rockstar Games, creators of the Grand Theft Auto franchise, said it was "very saddened" to hear of Mr Hall's death; 3- Oakland Athletics to Begin Accepting Bitcoin for Private Suites	