1 2

Tyche at Factify 2022: Fusion Networks for Multi-Modal Fact-Checking

Nainesh Hulke[1], Bharath Raj Siva[1], Ankesh Raj[1] and Ali Asgar Saifee[1]

[1]*Indian Institute of Technology (Indian School of Mines), Dhanbad, India*

Abstract

This paper describes our approach for the multimodal fact-checking (Factify) challenge at AAAI 2022. Fake news has the potential to harm people and society. To combat fake news, fact-checking becomes crucial. False claims are prevalent in both visual and textual forms. Multimodal techniques can merge information from both these modalities. In our approach, we treated this challenge as a multi-class classification task. We extracted textual and visual features from the texts and images, respectively. We used a single BERT module as a text feature extractor for both claim and reference text so that the extracted feature representations show the perspective of the same model on both claim and reference texts rather than two feature representations from two different models. The same goes for the image feature extractor. EfficientNet-B3 was used for extracting image features. Then the features were passed through a proposed fusion module and the classifier. The purpose of the fusion module is to enable the model to learn semantic similarities between texts and images. We achieved the best F1 score of 0.692 on the test set of the FACTIFY dataset.

Keywords

Multimodal Fact-Checking, Fusion Network, Fake News Detection

1. Introduction

News and information spread quickly through the internet and social media. This has also led to a vast increase in the spread of misinformation and fake news. As per a study, Facebook engagements with fake news sites average roughly 70 million per month[1]. Fake news has the potential to harm people and society. This spread has been blamed for incidents ranging from ethnic violence, inter-racial violence, and religious conflicts to mass riots. For example, some studies argue that fake news is gradually becoming an essential aspect of South Africa's xenophobia discourse[2].

There has been a lot of recent work on fact-checking claims. Early works on fact-checking focused on using textual information extracted from the text of the article, such as statistical text features[3], emotional information[4][5][6], or integrating metadata with the text[7]. Although

[1]All authors contributed equally.

[2]Team Name: Tyche

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

nainesh.18je0513@cse.iitism.ac.in (N. Hulke); bharathraj.18je0220@cse.iitism.ac.in (B. R. Siva); ankesh.18je0122@cse.iitism.ac.in (A. Raj); munni.18je0077@cse.iitism.ac.in (A. A. Saifee)

the textual content can be a significant indicator for fake news detection, it is not sufficient when used alone. Some researchers have proposed systems that use the credibility of the pages that post the news[8] or profile characteristics of the users that shared the post to detect the articles that contain manipulated content[9][10].

Online articles and posts usually contain more information in the form of images and social context that can be useful for fake news detection. Online news contains images that generally attract the attention of users. Images in fake and real news may follow different patterns or be modified to attract users' attention and make them share them. There are several reasons why an image may be deemed fake. In most cases, this involves digital manipulation, e.g., cropping, splicing, etc. However, there are cases when an image is entirely legitimate, but it is published alongside some text that does not reflect its content accurately.

Hence, it is essential that a system also exploits information extracted from the images for effective fake news detection. Visual information can complement the textual one for fake news detection. Multimodal systems can merge information from both these modalities. It is also capable of disambiguating the wrong classifications and improving the results using combined modalities. Some researchers have proposed multimodal systems that combine textual and visual information for determining whether an article or a post is fake or not[11][12] or combined textual, visual, and semantic information for fact-checking claims using a multimodal architecture[13].

2. Related Work

Early attempts to use deep learning methods to counter fake news employed data of a single modality. Vo et al. proposed a novel framework, Fact-checking Response Generator (FCRG)[14], to generate a fact-checking response tweet to combat fake news on Twitter. They showed distinguishing linguistic features of fact-checking tweets compared with Normal and Random Replies.

Wang et al.[7] introduced a new dataset LIAR for fake news detection and proposed a CNN model to integrate metadata with text. The proposed method outperforms text-only models strictly, suggesting the use of multiple modalities to express the news article's intent more clearly.

News articles often contain images and text; hence, multimodal deep learning methods have produced good results. Wang et al. [15] proposed the EANN (event adversarial neural network) model with three major components: the multimodal feature extractor, the fake news detector, and the event discriminator. The textual and visual latent feature representations are learned and concatenated to form the final multimodal feature representation.

Multimodal approaches that use features from the different modalities and feed them to different types of networks have also been successful. Gallo et al.[16] introduced multimodal fusion networks where the text and image feature representations are concatenated and passed through a neural network. The proposed method achieves the state-of-the-art results on the UPMC Food-101 dataset.

Recent research in multimodal fake news detection by Giachanou et al.[13] proposed a multi-modal multi-image network for fake news detection. The proposed system combines textual,

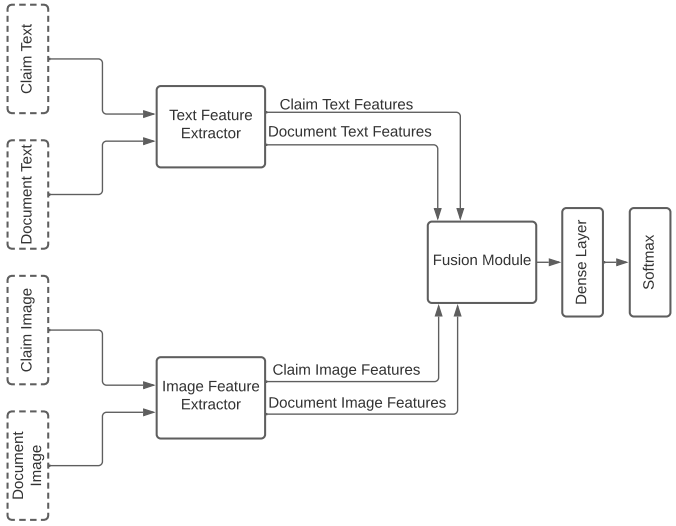

Figure 1: Model Architecture

visual, and semantic information. For the textual representation, BERT-Base [17] was used to capture the underlying semantic and contextual meaning. Image tags were extracted from multiple images that the articles contained using the VGG-16 model [18] for the visual representation. The semantic information was represented by the image-text similarity calculated using the title and image tags embeddings cosine similarity. All these features are concatenated to make the final prediction.

3. Proposed Approach

Our work is based on the FACTIFY dataset [19] [20]. In our approach, we extract the textual and visual features from the claim and reference texts and images, respectively. We propose a fusion module for the model to learn semantic similarities between the texts and images. The extracted features from this module are concatenated and passed through a classifier. Our approach aims to train a model that can use the relationship between the different modalities, i.e., text and image. Our model mainly consists of three different components Figure-1.

- Text Feature Extractor
- Image Feature Extractor
- Fusion Module

We used a single text feature extractor for both claim and reference text so that the extracted feature representations show the perspective of the same model on both the texts rather than two feature representations from two different models. A similar argument can be applied to images as well.

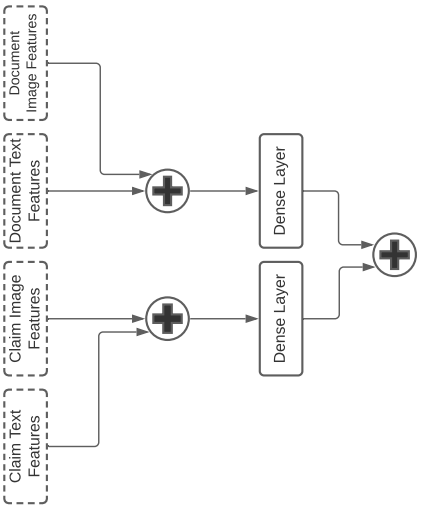

Figure 2: Text_Image_X2 Fusion Module

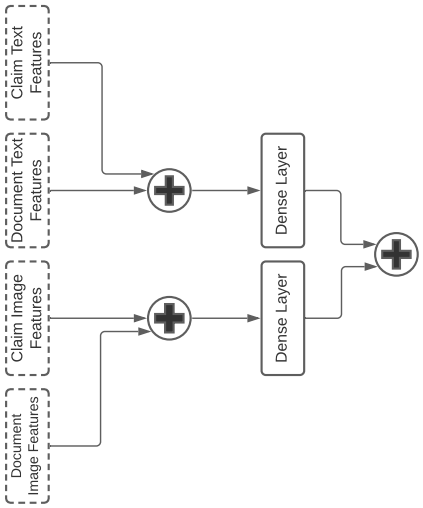

Figure 3: Text_X2_Image_X2 Fusion Module

3.1. Text Feature Extractor

We use the pre-trained Bidirectional Encoder Representations from Transformers (BERT) [17] base model because of its ability to tackle a broad set of NLP tasks. We fine-tuned the model for our specific task of text feature extraction. Specific details about the BERT model architecture, which we used, are given in Appendix A.

3.2. Image Feature Extractor

We wanted to have a model with high accuracy on ImageNet [21] that can be trainable with limited hardware resources, so we used the EfficientNetB3 architecture, pre-trained on ImageNet, and replaced the last dense layer with a dense layer of size 256 followed by a batch normalization layer. The EfficientNetB3 has 12 million parameters and achieves a top-1 accuracy of 81.6% on ImageNet. It balances the trade-off between classification accuracy and computational efficiency for our task.

3.3. Fusion Module

The extracted feature outputs are combined using a fusion module. We propose two methods of fusion with different intuitions on the output.

- **Text_Image_X2 :** In this method Figure-2., the claim text features and claim image features are concatenated and passed through a dense layer of size 256 to give the claim features. Similarly, we also get the reference features. This captures the similarity between text and image features of both claim and reference.

- **Text_X2_Image_X2 :** In this method Figure-3, the claim text features and reference text features are concatenated and passed through a dense layer of size 256 to give the text features. Similarly, we also get the image features. This captures the similarity between text features of claim and reference and image features of claim and reference.

The output features are then concatenated and passed through a dense layer of size 5. Finally, a Softmax layer is added to get the class probabilities.

4. Experiments

4.1. Data Preprocessing

The FACTIFY dataset has both text and images; therefore, preprocessing was performed separately.

4.1.1. Text Preprocessing

First, the text was converted to lowercase, removing all redundant elements, such as ASCII characters and weblinks. The stop words (i.e., articles, connectors, prepositions, and others) were removed since they do not contain any information and the semantics of sentences remain intact. We used the *nltk* [22] stopwords list to remove stopwords. The words of a sentence were then lemmatized. Lemmatization involves using a vocabulary and morphological analysis of words to remove inflectional endings only and return the base or dictionary form of the word. Thus, the size of the text was reduced, and the semantics of the sentence were intact. Only important words remained.

4.1.2. Image Preprocessing

The training data have a variable aspect ratio and size, so we adopted a stage-dependent re-scaling policy. We re-scale an image in the training stage so that the min side is 256 pixels wide while keeping the initial image's aspect ratio. As a result, the spatial information was not lost, and at the same time, the input to models was within trainable limits. Various other augments from Albumentations [23] were adopted to prevent overfitting. Since all the images are similar to real-world images, all arguments were considered to mirror real-world images.

4.2. Experimental Settings, Hyperparameter Tuning, and Results

We trained the text and image feature extractor together. The preprocessed images were resized to 256×256 before passing them through the image feature extractor model. We used the initial learning rate of 10^{-5} with a linear scheduler. For the BERT model, we used the initial learning rate of $3{\times}10^{-5}$ and kept decreasing it using the scheduler. The BERT sequence size was set to 64. A learning rate of 10^{-5} was used for the fusion module. F1 loss was used as the metric and Adam[24] as the optimizer.

The best way to maximize the F1 metric would be to minimize the F1 loss, defined as the 1−F1 score. The issue is that the F1 score is not differentiable. We took inspiration from Lee et

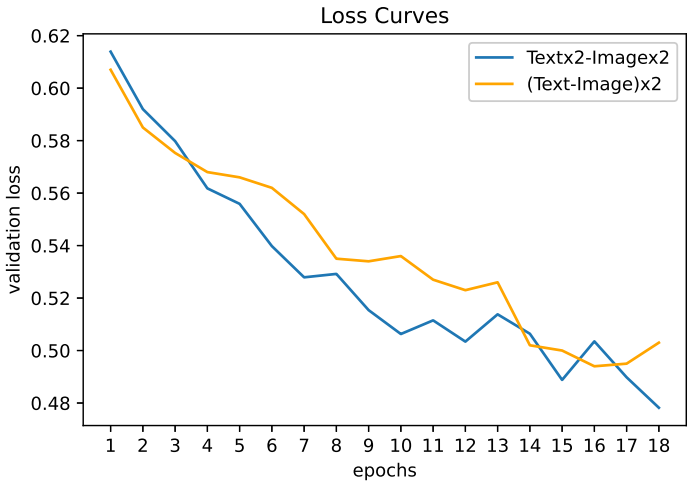

Figure 4: Training Curves for proposed Fusion Modules

Table 1
F1 Score analysis for different models

Model Name	Validation F1 score
Bert-base	0.673
EfficientNet-B3	0.617
EfficientNet-B3 + Bert-base (Early Fusion)	0.689
EfficientNet-B3 + Bert-base (Text_X2_Image_X2)	0.725
EfficientNet-B3 + Bert-base (Text_Image_X2)	0.705
Ensemble of Text_X2_Image_X2 and Text_Image_X2	0.712

al.[25] and implemented it for multiclass classification. We accept probabilities instead of actual counts of true positive, true negative, false positive, or false negative. Say the class ”refute” is predicted with probability 0.2, while the true label is ”refute”. Then we calculate true positive as 0.2 and false negative as 0.8.

We trained the model using the NVIDIA Tesla P100 GPU with 16 GB of memory and also used RTX5000, having 32 GB of memory.

Text_X2_Image_X2 fusion module was trained for 18 epochs, and the model with Text_Image_X2 fusion module was trained for 17 epochs, after which the decrease in validation loss Figure-4 became insignificant.

We also tried different architectures, as mentioned in Table 1. We observed that the text-only classifier performs better than the image-only classifier indicating that text provides better insights to news articles. We trained on full data (i.e., text and image combined) using EfficientNetB3 for image and BERT for text which is then followed by an early fusion module [16](a fusion module that concatenates all the extracted features together and passes them through a Dense layer). Then we tried our proposed models Text_X2_Image_X2 and

Text_Image_X2 and found that these outperformed both the early fusion models and the models trained on a single modality.

The Text_X2_Image_X2 model gave an F1 score of 0.679 on the test set performing better than the Text_Image_X2 model, which gave a score of 0.664. This was due to the way we concatenate the extracted features. Semantic similarities between similar data(text to text and image to image) lead to a better estimation of differences between claim and reference. We ensembled the models doing a weighted average with different ratios and found an increase in the F1 score to 0.692 on the test set (equal weights).

5. Conclusion and Future Work

Combating fake news is crucial as it can destroy one's credibility and hurt many people. In this paper, we focused on countering the problem of fake news using fact-checking with the help of the FACTIFY dataset. We proposed two novel methods for combining the multimodal features. The Text_X2_Image_X2 method leverages the similarity between the claim and document text features and claim and document image features to predict the required class. The Text_Image_X2 method leverages the similarity between the text and image claim features and text and image document features to predict the required class. Both the proposed methods outperform the existing early fusion network. Our proposed methods showed that combining similar features separately(EfficientNetB3 + Bert-base + Text_X2_Image_X2 and EfficientNetB3 + Bert-base + Text_Image_X2) is more effective than combining all the features together(EfficientNetB3 + Bert-base + Early Fusion). Future work expects that using deeper CNN models for image feature extractors can improve model performance. Similarly, we can use different variants of BERT for a more robust text feature extractor. Playing with the neural networks used in fusion modules and classifiers can also give better results.

6. Acknowledgement

We are indebted to team CyberLabs for their generous support. Also, our work was made possible with the support of JarvisLabs.ai. We thank them for providing a cloud GPU instance for our experiments. We also thank the organisers of DE-FACTIFY 2022 for giving us the opportunity to work on the dataset.

References

[1] H. Allcott, M. Gentzkow, C. Yu, Trends in the diffusion of misinformation on social media, Research & Politics 6 (2019) 2053168019848554.

[2] V. Chenzi, Fake news, social media and xenophobia in south africa, African Identities (2020) 1–20.

[3] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: Proceedings of the 20th international conference on World wide web, 2011, pp. 675–684.

[4] O. Ajao, D. Bhowmik, S. Zargari, Sentiment aware fake news detection on online social networks, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2507–2511.
[5] B. Ghanem, P. Rosso, F. Rangel, An emotional analysis of false information in social media and news articles, ACM Transactions on Internet Technology (TOIT) 20 (2020) 1–18.
[6] A. Giachanou, P. Rosso, F. Crestani, Leveraging emotional signals for credibility detection, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 877–880.
[7] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
[8] K. Popat, S. Mukherjee, A. Yates, G. Weikum, Declare: Debunking fake news and false claims using evidence-aware deep learning, arXiv preprint arXiv:1809.06416 (2018).
[9] A. Giachanou, E. A. Ríssola, B. Ghanem, F. Crestani, P. Rosso, The role of personality and linguistic patterns in discriminating between fake news spreaders and fact checkers, in: International Conference on Applications of Natural Language to Information Systems, Springer, 2020, pp. 181–192.
[10] K. Shu, S. Wang, H. Liu, Understanding user profiles on social media for fake news detection, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), IEEE, 2018, pp. 430–435.
[11] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, S. Satoh, Spotfake: A multi-modal framework for fake news detection, in: 2019 IEEE fifth international conference on multimedia big data (BigMM), IEEE, 2019, pp. 39–47.
[12] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.
[13] A. Giachanou, G. Zhang, P. Rosso, Multimodal multi-image fake news detection, 2020, pp. 647–654. doi:10.1109/DSAA49011.2020.00091.
[14] N. Vo, K. Lee, Learning from fact-checkers: Analysis and generation of fact-checking language, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 335–344.
[15] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, 2018, pp. 849–857. doi:10.1145/3219819.3219903.
[16] I. Gallo, G. Ria, N. Landro, R. L. Grassa, Image and text fusion for upmc food-101 using bert and cnns, in: 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), 2020, pp. 1–6. doi:10.1109/IVCNZ51579.2020.9290622.
[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[18] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[19] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Factify: A multi-modal fact verification dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.

[20] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Benchmarking multi-modal entailment for fact verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.
[22] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, " O'Reilly Media, Inc.", 2009.
[23] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A. A. Kalinin, Albumentations: fast and flexible image augmentations, Information 11 (2020) 125.
[24] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[25] N. Lee, H. Yang, H. Yoo, A surrogate loss function for optimization of F_β score in binary classification with imbalanced data, arXiv preprint arXiv:2104.01459 (2021).

A. Text Feature Extractor Details

The model used was the BASE version of the BERT model having the following parameters: English language uncased, 12 hidden layers (L), 768 hidden sizes (H), 12 self-attention heads (A), 30522 words dictionary (vocab size), 110 million parameters in total. Additionally, a Dropout layer with a probability of 0.3, a Dense layer of size 256, and a Batch-Normalization layer are added to the final layer corresponding to the CLS token.