Team coco at Factify 2: Optimizing Modality Representation and Multimodal Fusion for Multimodal Fact Verification

Kangshuai Guo[1], Shichao Luo[3], Ruipeng Ma[1], Yan Wang[1] and Yanru Zhang[1,2,*]

[1]University of Electronic Science and Technology of China
[2]Shenzhen Institute for Advanced Study of UESTC
[3]CITIC aiBank Corporation Limited China

Abstract

While social media has changed news dissemination and how humans obtain information, it has also become the main channel for disseminating fake news. Quickly identifying fake news in social media and curbing the spread of false information is crucial to purifying cyberspace and maintaining public safety. Exploring efficient modality representation and multimodal information fusion methods has been a hot topic in the field of multimodal fake news detection or fact verification. To this end, a new multi-Modal fact verification is proposed: First, deep modality representations of text and images are extracted using a large-scale pre-trained model. Secondly, a bidirectional-hybrid attention mechanism is introduced to fuse text and image features. The hybrid mechanism reduces redundant information generated during multimodal fusion and uses Bidirectional feature fusion to ensure the integrity of information. Besides, we adopt the ensemble method to achieve better performance. Our team, coco, won the sixth prize (F1-score: 75.696%) in the Factify challenge hosted by De-Factify @ AAAI 2023. Extensive experiments including comparison experiments, analysis of parameter sensitivity, and ablation study demonstrate the effectiveness of our proposed approach.

Keywords

Multi-modal fact verification, Modality representation, Multi-modal fusion, De-Factify

1. Introduction

Due to the timeliness and profitability of social media, some people artificially fabricate or adapt news to generate fake news in order to obtain attention traffic. Fake news can be easily disseminated along with real news, thereby confusing the majority of users, which has brought certain harm to the economy and society. Especially since the COVID-19 epidemic, a large amount of fake news has emerged on social media, which has grown exponentially in a short period of time, bringing negative and negative impacts on society [1].

The current form of news is no longer limited to text, but a combination of text, image, video and other modalities. Compared with traditional single-modal text news, multi-modal news is

De-Factify 2: 2nd Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2023. 2023 Washington, DC, USA

*Corresponding author.
guokangshuai@gmail.com (K. Guo); xiaoluoyfy@gmail.com (S. Luo); 202221080135@std.uestc.edu.cn (R. Ma); yanbo1990@uestc.edu.cn (Y. Wang); yanruzhang@uestc.edu.cn (Y. Zhang)

easier to attract people's attention. Fake news usually has highly emotionally provocative text and visually impactful pictures or videos. In addition, due to the mixture of real information and fake news, fake news is generally difficult to be identified by humans. Therefore multimodal fact verification is one of the effective ways to combat fake news. Multimodal fact verification is defined as: discriminating a given multimodal claim (text + visual)as true/false given credible news sources.

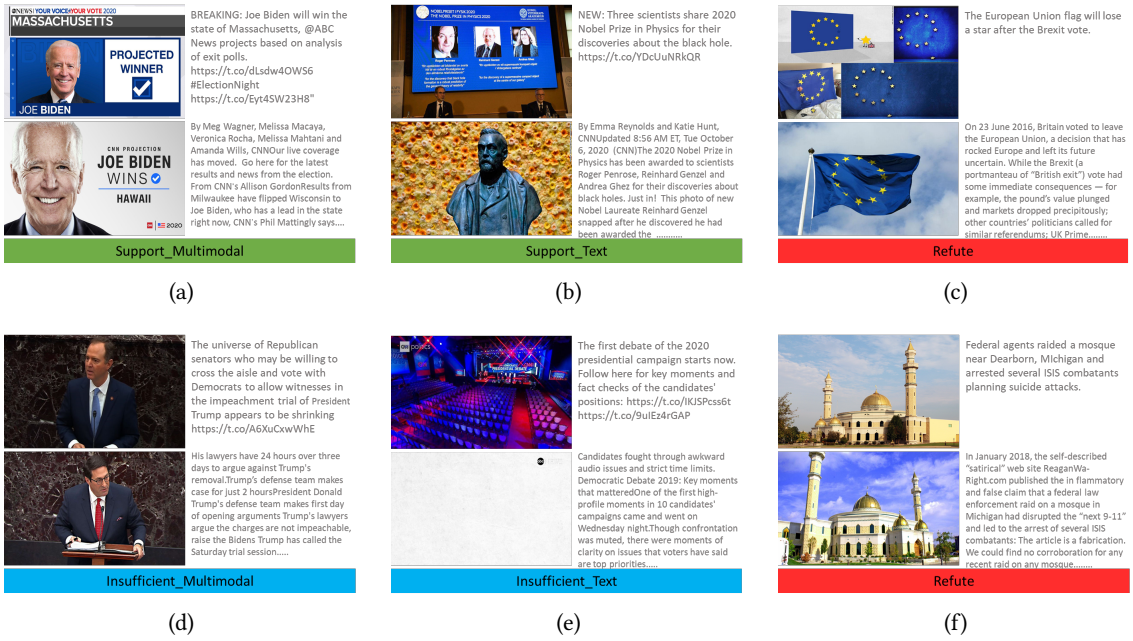


Figure 1: Visualization of examples from the Factify 2 dataset [2], that sample examples for all five categories.

The challenge Factify_2 [1] hosted by De-Factify team[2] provides a more complex and efficient multimodal fact verification task than just classifying claims as true or false [2, 3, 4]. The goal is to design a method to classify the given claim text and images into one of the five categories: Support_Multimodal, Support_Text, Insufficient_Multimodal, Insufficient_Text, and Refute. Fig. 1 shows some examples for all five categories.

To tackle the task, in this paper, we propose a new multi-Modal fact verification method, which by optimizing modality representation and multimodal fusion to Improve multimodal fact verification accuracy. Specifically, deep modality representations of text and images are extracted using a large-scale pre-trained model which is combined with a positionally encoded self-attention mechanism. Afterward, a bidirectional hybrid attention mechanism is introduced to achieve homomodal and cross-modal information fusion. The hybrid mechanism reduces redundant information generated during multimodal fusion and uses Bidirectional feature fusion to ensure the integrity of information. Finally, we adopt the ensemble method to achieve

[1]https://codalab.lisn.upsaclay.fr/competitions/8275
[2]https://aiisc.ai/defactify2/index.html

better performance.

The main results of this paper can be summarized as follows:

- First, we explore and compare different pre-trained models and embedding methods for modality representation, enabling efficient fact verification.
- Second, we improve detection performance by building embedding pairs of modality representations modeling to achieve multimodal alignment relationships and performing multimodal information fusion.
- Finally, improve the effectiveness and generalization of the model through ensemble learning.

The rest of the paper is organized as follows: in the next section, we first review the previous studies on the multimodal fact checking and pre-trained models. Then, the details of the framework of the proposed method is introduced in section III. After that, experimental results and analysis will be explained in Section IV. And finally, we would give our conclusion in Section V.

2. Related Works

2.1. Multimodal Fact Checking

Existing works on multimodal fake news detection or fact-checking are reviewed briefly. Early studies [5, 6, 7, 8] on fake news detection or fact-checking were based only on unimodal information (text content), and these research methods can be summarized as feature-building techniques and deep learning techniques. The current form of news is no longer limited to text, but is composed of multiple modalities such as text, images, and videos. Recent studies about fake news detection or fact-checking have started to take images [9, 10, 11, 12] and videos [13] into consideration. Compared with single-modal fake news detection techniques, multi-modal fake news is more flexible, authentic and accurate. Numerous studies [14, 15, 16] have shown that multimodal fake news detection models perform better than single modality models under the same dataset. Most approaches for multimodal fake news detection or fact-checking are based on cross-modality consistency checking [17, 18] or fusion of multi-modal (text + visual) information by modeling multi-modal alignment relationships [19, 14, 20]. The former focuses on multimodal consistency measurement [21]. SAFE [22] uses an Image Captioning model to translate images into sentences, and then computes multimodal inconsistency by measuring the sentence similarity between the original news text and the generated image captions. MCNN [23] transforms text and visual features into a common feature space to calculate similarity through sub-network weight sharing. The latter improves detection performance by computing fused representation of multimodal information [21]. attRNN [19] proposes a recurrent neural network based on a neuron-level attention mechanism to fuse graphic information. CARMN [24] uses a collaborative attention mechanism to model bidirectional augmentation between text and images. EMAF [25] extracts the target labels of the images, and then uses the capsule network to fuse the nouns in the text with these target labels. Our approach focuses on multimodal consistency measures, transforming textual and visual features into a common feature space to compute similarity, which is inspired by SAFE and MCNN.

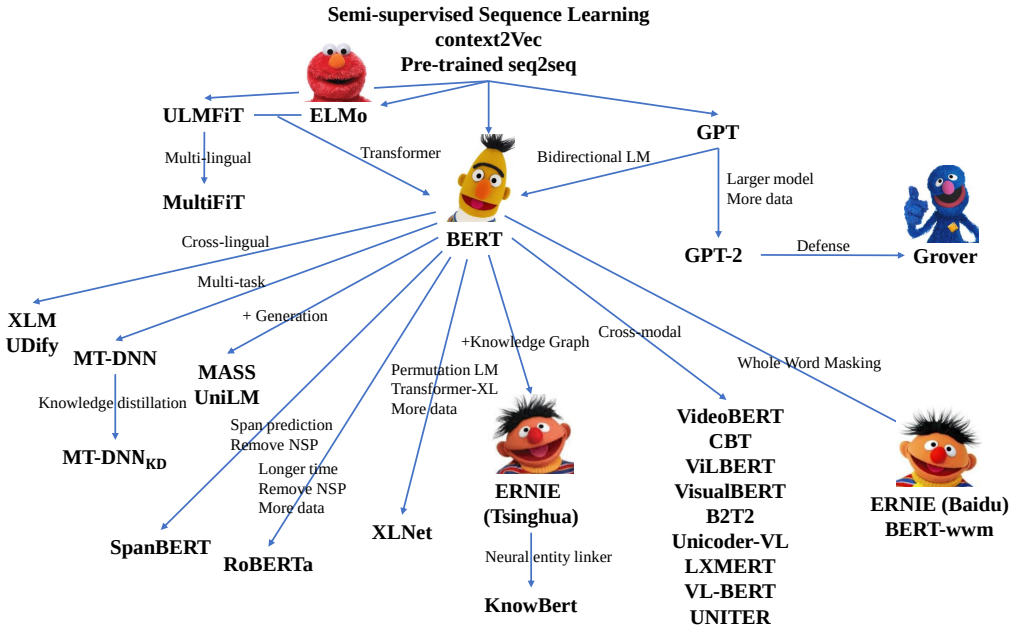


Figure 2: The family of recent typical PTMs, including both pre-trained language models and multimodal models. [26].

2.2. Pre-trained Models

Transformer [27] became the preferred architecture for language models in 2017, followed by the emergence of GPT [28] and BERT[29] in 2018, bringing Pre-Trained Models(PTM) into a new era [26]. Common PTMs are shown in Fig. 4. These PTMs models are very large, with a large number of parameters, which can capture information such as polysemy, morphology, syntactic structure, and real-world knowledge from the text, and then fine-tune the model to achieve amazing performance on downstream tasks [30]. By now, fine-tuning for specific tasks on large-scale PTMs has become an industry consensus. The rise of a series of large-scale PLMs such as DeBERTa [31], Roberta [32], XLM-RoBERTa [33], XLNet [34], and SpanBERT [35]. These PLMs have been fine-tuned using a few label examples with task-specific and have created a new state of the art in many downstream tasks. [36].

In the past 10 years, Convolutional Neural Network (CNN) [37], as a model that is good at capturing local features, has been placed high hopes in the field of computer vision and has led an era. However, the operation of convolution lacks a global understanding of the image itself, and cannot model the dependencies between features, so it cannot fully utilize contextual information. Vision Transformer [38] is therefore proposed, an image classification method based entirely on the self-attention mechanism. Compared with CNN, the Transformer architecture has achieved good results in many visual tasks since its self-attention mechanism is not limited by local interaction features, it can mine long-distance dependencies and learn the most appropriate inductive bias according to different task objectives [39]. Then the rise of a series of large-scale transformers architecture visual pre-training models such as ViT [38], Deit [40], Swin Transformer [41], and BEiT [42].

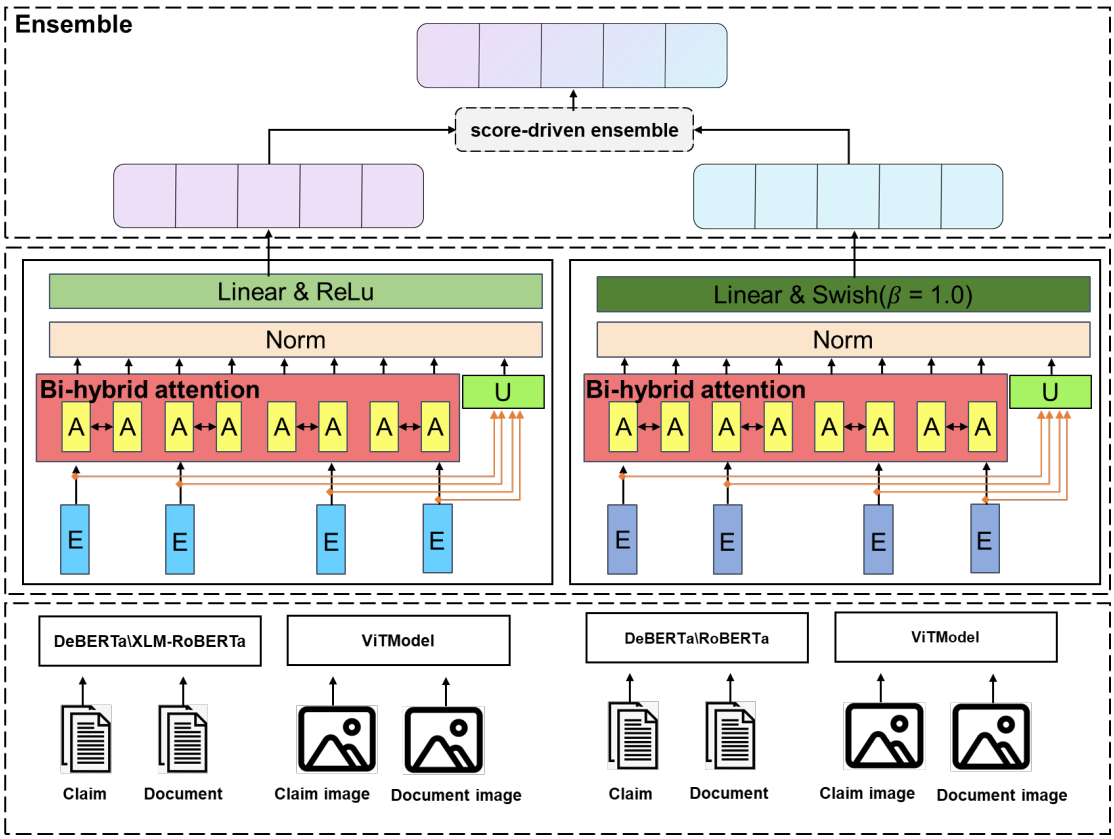


Figure 3: Illustration of the method framework.Our proposed method is ensembled by two different models with similar architecture.

3. Method

3.1. Overview

Figure 3 illustrates the overview of the proposed framework. The input of each model contains the claim text, the document text,the claim image, and the document image. The modality representation part adopts DeBERTa as the pre-trained NLP model and DeiT as the pre-trained CV model and feeds the outputs of pre-trained models to the embedding layer for transforming modality representation into corresponding embeddings, which to help modal information alignment. The multi-modality fusion part fuses this information from the homomodal (text pair, image pair) and cross-modal (text-image pair, image-text pair) based on bidirectional-hybrid attention mechanism. The classifier predicts the probability of each category based on the embeddings from modality representation and the embeddings from multi-modality fusion. Finally, the output of each model is ensembled according to the score-driven strategy to improve fact-checking effectiveness and generalization.

3.2. Modality Representation

3.2.1. Pre-training

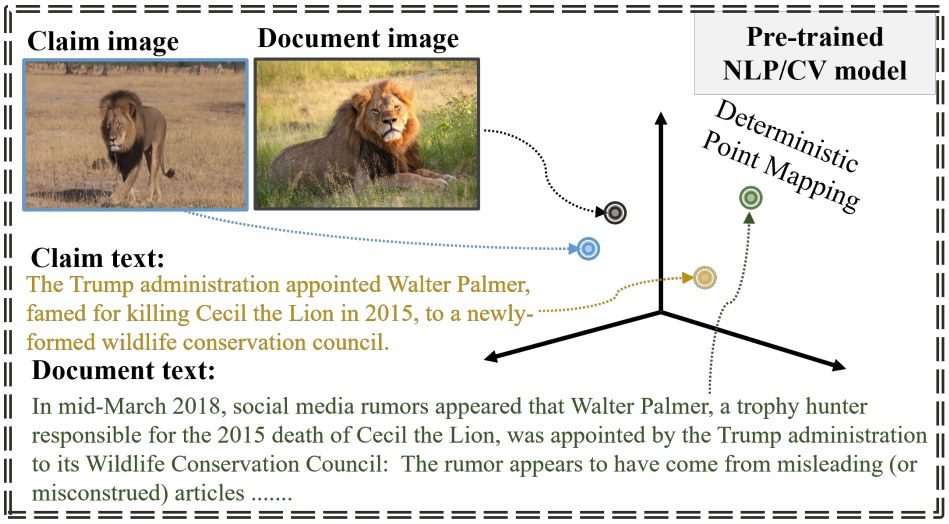


Figure 4: Illustration of modality representation via pre-training models. Obtain deterministic point mapping of different modalities via pre-trained cv/nlp models.

Pre-training obtains task-independent pre-training models from large-scale data through self-supervised learning. The exploration of pre-trained models is mainly devoted to deep semantic representation and contextual semantic representation. Extensive work [26, 30] has shown that pre-trained models on large corpora can learn general-purpose language representations, avoiding training new models from scratch when solving downstream tasks. Compared with other pre-training models, Deberta implements Disentangled Attention and Disentangled Attention. The former enables it to have stronger representation capabilities, and the latter is used to avoid inconsistencies between pre-training tasks and downstream tasks. To this end, we use DeBERTa as our pre-trained NLP model and DeiT as our pre-trained CV model to modality representation. The abstraction of modality representation is shown in Fig.4.

3.2.2. Embedding

In order to adapt to downstream tasks, the output of the pre-trained model is fed into the embedding layer. As shown in Fig. 5. The text and image modality representations are mapped to a unified embedding space, which enhances the alignment relationship between multimodality including homomodal and cross-modality, and is conducive to multi-modal fusion.

3.3. Multi-Modality Fusion

For the homomodal, consider their mutual relationship by performing uni-modality (text or image) similarity matching, only need to pay attention to the common modality representation. For cross-modality, only part of the information contained in the image is related to the text,

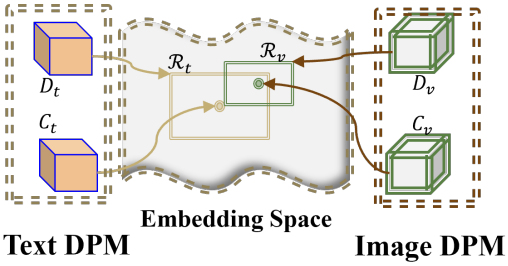
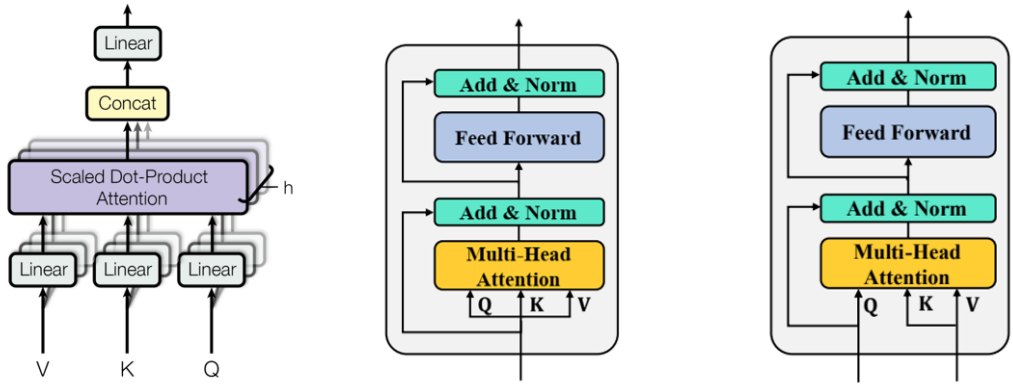


Figure 5: Illustration of the embedding space. Mapping DPMs of different modalities to the unified embedding space [43].

and there is a large amount of redundant information irrelevant to the task. so the information fusion of same-modality and cross-modal is necessary.

The multi-head attention mechanism endows the model with the ability to jointly attend to information from different representation subspaces at different positions [27]. The co-attention mechanism [44] is a variant of the standard multi-head self-attention mechanism which contributes to multimodal information fusion. Their structure is shown in Fig. 6



(a) multi-head attention [27] (b) multi-head self-attention block[44] (c) Co-attention block[44]

Figure 6: Structural diagrams of several common attention mechanisms

Our Bidirectional-hybrid attention mechanism considers both the same-modality and cross-modality in multimodal fusion. The hybrid mechanism reduces redundant information generated during multimodal fusion and uses bidirectional feature fusion to ensure the integrity of information. For the same modality, we construct text pairs and image pairs based on the output of the embedding layer. Correspondingly, cross-modality constructs text-image pairs and image-text pairs. Then multi-modal fusion is achieved by adopting the co-attention block (Fig.6c).

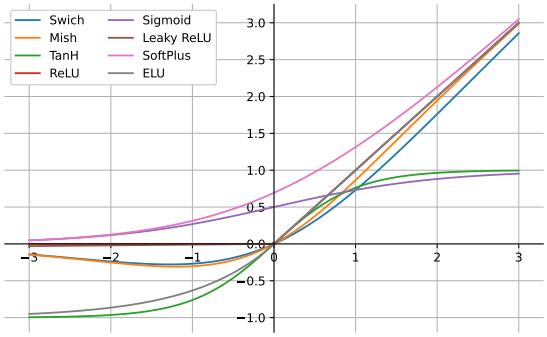


Figure 7: Common Activation Functions.

3.4. Classifier & Ensemble Method

3.4.1. Classifier

The multimodal fact verification task is a multi-classification problem, specifically, subdivided into 5 categories: Support_multimodal , Support_textual , Insufficient_multimodal , Insufficient_text and Refute.

The embedding of modality representation undergoes multi-modal fusion through a bidirectional hybrid attention mechanism. The output passes through a regularization layer and then enters a classifier, which consists of the linear layer and an activation function. The activation function realizes the effect of nonlinear transformation. Common activation functions are shown in Fig. 7. In deep neural networks, it is crucial to use a suitable mapping for nonlinear fitting to complete the classification task [45]. Our approach put attention on selection of activation functions, specifically, two activation functions, ReLU and Swish ($beta$ = 1), are used in the classifier and embedding layers of the model to achieve efficient multimodal fact-checking effects.

3.4.2. Ensemble

The hybrid prediction method combines different prediction methods to improve the performance of the final prediction. The performance of a single model is limited in many cases. Hybrid forecasting methods combine the ability of multiple models to accommodate changes in the sample by setting the results of each model in combination [46].

As shown in Fig. 3. The final output of our method is ensembled by two different multi-class prediction models. We adopt a score-driven ensemble strategy, where weights are determined based on the Val-set f1 score of each model, in order to balance the effect of each model and achieve better multimodal fact-checking results.

4. Experiment

4.1. Dataset

There are many opensource datasets in the field of automated factchecking, such as LIAR [47], FEVER [48] , Covid19 Fake News dataset [49] and Claim matching beyond english [50]. Compared with the Factify 2 datasets [2], which contain complete images and textual information about claims and reference documents, the aforementioned are all unimodal datasets (only text). Each sample of includes claim, claim_image, claim_ocr, document, document_image, document_ocr, and category.

The description for each attribute is as follows [4]:

- claim: the text of the claim source by tweet A.
- claim_image: the image of tweet A.
- claim_ocr: ocr of claim image.
- document: the article text of the given reference document which is tweet B.
- document_image: the image of tweet B.
- document_ocr: ocr of document image.
- category: a given category of all five classes..

The category including:

- Support_Multimodal: text and images of claim are supported by the document corresponding.
- Support_Text: claim text is supported by the document text, but claim images are not relevant.
- Insufficient_Multimodal: claim text is neither supported nor refuted by the document text but images are similar to the document images.
- Insufficient_Text: both text and images of the claim are neither supported nor refuted by the document corresponding.
- Refute: fake claim or fake image inferred from document corresponding.

The training set [3] contains 35,000 samples, which has 7,000 samples of each category, and the validation set contains 7,500 samples, which has 1,500 samples of each category. The test set, which is used to evaluate the leaderboard score, same specifications as the validation set [4, 36].

4.2. Results

4.2.1. Testing Performance

Table 1 shows the performance of the testing set. Our approach achieved 0.75696 of the weighted average F1-score, winning the sixth prize in multi-modal fact checking. This result outperformed the baseline by 11%. Compared with the Pre-CoFact method, our method pays more attention to the prediction of the support category, and the category of insufficient is slightly weaker.

[3] https://drive.google.com/drive/folders/13JwnIBzDfe8a5E1anPkt7J90r4NBIYES

Table 1
Performance of our model in terms of testing score. Our method achieved sixth prize. *Improv* row shows the absolute improvements of the ensemble model over Pre-CoFact model.

	Support		Insufficient		Refute	Total
	Text	Multimodal	Text	Multimodal		
Deberta_Relu	0.78542	0.82995	0.76186	0.79807	0.99833	0.83473
Deberta_Swish	0.76339	0.84748	0.81422	0.81750	0.99899	0.84832
Roberta_Relu	0.74108	0.85627	0.77750	0.80905	0.99966	0.83671
Roberta_Swish	0.77949	0.82537	0.81563	0.79129	0.99631	0.84162
XLM-Roberta	0.75363	0.84214	0.74999	0.79032	0.99866	0.82695
Ensemble model	0.77250	**0.86492**	0.81516	**0.82995**	**1**	**0.85651**
Baseline	0.5	0.82720	0.80239	0.75930	0.98819	0.64990
Pre-CoFact [36]	0.68881	0.81610	0.84836	0.88309	1	0.74585 [4]
Our model	**0.77250**	**0.86492**	0.81516	0.82995	**1**	**0.75696** [5]
Improv	*+0.08369*	*+0.04882*	*-0.03320*	*-0.05314*	-	*+0.01111*

4.3. Further Analysis

4.3.1. Comparison Result

In order to show that different pre-training models affect the modality representation, we did some comparative experiments. As shown in Fig. 8. In modality representation, using different pre-trained models has a certain impact on the performance of fact-checking. It can be seen that using the Deberta pre-trained model for modal representation has achieved better results.

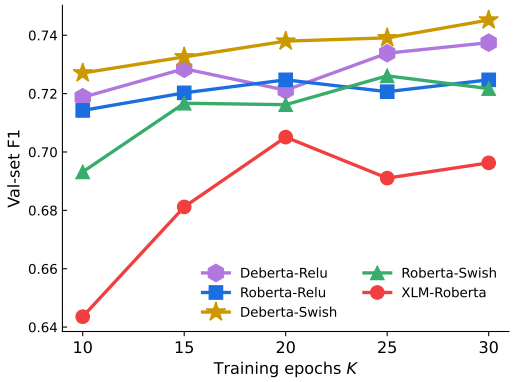


Figure 8: Performance vs PLM Varying

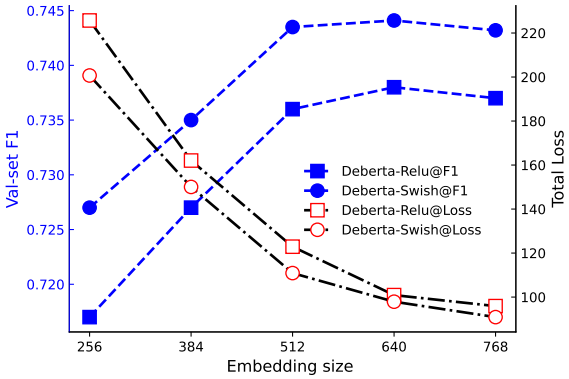


Figure 9: Performance vs Embedding Size Varying

4.3.2. Parameter Sensitivity

The presence of embedding layers helps modality representations to better adapt to downstream tasks. Mapping different modality representations into a unified embedding space reduce the

variance between modalities, which facilitates modality representation feature alignment and cross-modal fusion. Parameter sensitivity experiments on embedding size are performed, which explore the effect of embedding space size on fact-checking performance. As shown in Fig. 9.

4.3.3. Ablation Study

Table 2
Ablation study of different modules of our method.

Method	Loss_{avg}	Val_set F1	Test_set F1
Deberta_Relu	0.03154	0.73528	0.83473
w/o attention	0.40159	0.71418	0.82958
Deberta_Swish	0.10254	0.74937	0.84832
w/o attention	0.55805	0.69580	0.81022
Ensemble model	-	-	0.85650
w/o ϕ	-	-	0.85098
w/o avg	-	-	0.85323
w/o attention	-	-	0.82095

To evaluate the effect of the key components of our proposed model, we conduct an ablation study on these components based on the dataset. The ablation settings are as follows:

- w/o attention: Each model removes the bidirectional-hybrid attention mechanism during training, and only uses the embedding of modality representation.
- w/o ϕ : Model ensemble without the score-driven strategy, the same weight for each model.
- w/o avg: Model ensemble strategy without averaging model predictions.

The results are shown in Table 2. From this table, it is observed that all the components are indispensable for the superior performance of our method.

5. Conclusion

In this paper, we optimize modality representation and multimodal fusion achieve the efficient multimodal fact-checking task. Specifically, deep modality representations of text and images are extracted using a large-scale pre-trained model. Afterward, a bidirectional hybrid attention mechanism is introduced to achieve homomodal and cross-modal information fusion. To achieve better performance, we adopted an ensemble method by weighting several models. Extensive experiments including comparison experiments, analysis of parameter sensitivity, and ablation study demonstrate the effectiveness of our proposed approach.

[4]Data from the corresponding paper, also the official results De-Factify '22.
[5]The final score is obtained by a weighted average of the category-wise scores by the organizers.

Acknowledgement

We appreciate previous work [4, 51, 43, 52] and open resources[44, 36]. Based on this, we conducted our work on Defactify 2. We also appreciate the help from the Defactify team and program chairs. We appreciate Wu and Wang for their provided open resource.

References

[1] M. S. Islam, T. Sarkar, S. H. Khan, A.-H. M. Kamal, S. M. Hasan, A. Kabir, D. Yeasmin, M. A. Islam, K. I. A. Chowdhury, K. S. Anwar, et al., Covid-19–related infodemic and its impact on public health: A global social media analysis, The American journal of tropical medicine and hygiene 103 (2020) 1621.

[2] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. Reganti, A. Chadha, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Factify 2: A multimodal fake news and satire news dataset, in: proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.

[3] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, A. Das, A. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Findings of factify 2: multimodal fake news detection, in: proceedings of defactify 2: second workshop on Multimodal Fact-Checking and Hate Speech Detection, CEUR, 2023.

[4] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, et al., Factify: A multi-modal fact verification dataset, in: Proceedings of the First Workshop on Multimodal Fact-Checking and Hate Speech Detection (DE-FACTIFY), 2022.

[5] A. Hanselowski, C. Stab, C. Schulz, Z. Li, I. Gurevych, A richly annotated corpus for different tasks in automated fact-checking, arXiv preprint arXiv:1911.01214 (2019).

[6] N. Kotonya, F. Toni, Explainable automated fact-checking for public health claims, arXiv preprint arXiv:2010.09926 (2020).

[7] I. Augenstein, C. Lioma, D. Wang, L. C. Lima, C. Hansen, C. Hansen, J. G. Simonsen, Multifc: A real-world multi-domain dataset for evidence-based fact checking of claims, arXiv preprint arXiv:1909.03242 (2019).

[8] J. Nørregaard, B. D. Horne, S. Adalı, Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles, in: Proceedings of the international AAAI conference on web and social media, volume 13, 2019, pp. 630–638.

[9] C. Boididou, K. Andreadou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, Y. Kompatsiaris, et al., Verifying multimedia use at mediaeval 2015., MediaEval 3 (2015) 7.

[10] K. Nakamura, S. Levy, W. Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, arXiv preprint arXiv:1911.03854 (2019).

[11] S. Jindal, R. Sood, R. Singh, M. Vatsa, T. Chakraborty, Newsbag: a multi-modal benchmark dataset for fake news detection, in: CEUR Workshop Proc., volume 2560, 2020, pp. 138–145.

[12] J. C. Reis, P. Melo, K. Garimella, J. M. Almeida, D. Eckles, F. Benevenuto, A dataset of fact-checked images shared on whatsapp during the brazilian and indian elections, in:

Proceedings of the International AAAI Conference on Web and Social Media, volume 14, 2020, pp. 903–908.
[13] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, I. Kompatsiaris, A corpus of debunked and verified user-generated videos, Online information review 43 (2018) 72–88.
[14] 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.
[15] S. Qian, J. Hu, Q. Fang, C. Xu, Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection, ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 17 (2021) 1–23.
[16] M. Mayank, S. Sharma, R. Sharma, Deap-faked: knowledge graph based approach for fake news detection, arXiv preprint arXiv:2107.10648 (2021).
[17] Y. Wang, F. Ma, H. Wang, K. Jha, J. Gao, Multimodal emergent fake news detection via meta neural process networks, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3708–3716.
[18] S. Abdelnabi, R. Hasan, M. 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.
[19] Z. Jin, J. Cao, H. Guo, Y. Zhang, J. Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 795–816.
[20] J. Wang, H. Mao, H. Li, Fmfn: Fine-grained multimodal fusion networks for fake news detection, Applied Sciences 12 (2022) 1093.
[21] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, Y. Yu, Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1212–1220.
[22] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news detection, The World Wide Web Conference (2019).
[23] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, L. Wei, Detecting fake news by exploring the consistency of multimodal data, Information Processing & Management 58 (2021) 102610.
[24] C. Song, N. Ning, Y. Zhang, B. Wu, A multimodal fake news detection model based on cross-modal attention residual and multichannel convolutional neural networks, Information Processing & Management 58 (2021) 102437.
[25] P. Li, X. Sun, H. Yu, Y. Tian, F. Yao, G. Xu, Entity-oriented multi-modal alignment and fusion network for fake news detection, IEEE Transactions on Multimedia (2021).
[26] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, et al., Pre-trained models: Past, present and future, AI Open 2 (2021) 225–250.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[28] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).
[29] 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).
[30] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Science China Technological Sciences 63 (2020) 1872–1897.
[31] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).
[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[33] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, arXiv preprint arXiv:1911.02116 (2019).
[34] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, Advances in neural information processing systems 32 (2019).
[35] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, Spanbert: Improving pre-training by representing and predicting spans, Transactions of the Association for Computational Linguistics 8 (2020) 64–77.
[36] W.-Y. Wang, W.-C. Peng, Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification, arXiv preprint arXiv:2201.11664 (2022).
[37] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (1989) 541–551.
[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).
[39] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., A survey on vision transformer, IEEE transactions on pattern analysis and machine intelligence (2022).
[40] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.
[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[42] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).
[43] Z. Wang, Z. Gao, X. Xu, Y. Luo, Y. Yang, H. T. Shen, Point to rectangle matching for image text retrieval, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4977–4986.
[44] Y. Wu, P. Zhan, Y. Zhang, L. Wang, Z. Xu, Multimodal fusion with co-attention networks for fake news detection, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, 2021, pp. 2560–2569.
[45] D. Misra, Mish: A self regularized non-monotonic neural activation function, arXiv

preprint arXiv:1908.08681 4 (2019) 10–48550.
[46] H. Liu, trymore: Solution to spatial dynamic wind power forecasting for kdd cup 2022 (2022).
[47] W. Y. Wang, " liar, liar pants on fire": A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
[48] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).
[49] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, T. Chakraborty, Fighting an infodemic: Covid-19 fake news dataset, in: International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer, 2021, pp. 21–29.
[50] A. Kazemi, K. Garimella, D. Gaffney, S. A. Hale, Claim matching beyond english to scale global fact-checking, arXiv preprint arXiv:2106.00853 (2021).
[51] 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.
[52] L. Gao, Q. Zhang, X. Zhu, J. Song, H. T. Shen, Staircase sign method for boosting adversarial attacks, arXiv preprint arXiv:2104.09722 (2021).