EasyChair Preprint

№ 10104

Team Zhang at Factify 2: Unimodal Feature-Enhanced and Cross-Modal Correlation Learning for Multi-Modal Fact Verification

Fanrui Zhang, Qiang Zhang, Jiawei Liu and Esther Sun

May 12, 2023

Team Zhang at Factify 2: Unimodal Feature-enhanced and Cross-modal Correlation learning for Multi-Modal Fact Verification

Fanrui Zhang[1], Qiang Zhang[1], Jiawei Liu[1] and Esther Sun[2]

[1]University of Science and Technology of China
[2]University of Toronto

Abstract

In recent years, social media has enabled users to get exposed to a myriad of misinformation and disinformation which have attracted a great deal of attention in research fields. Despite the progress in text-based fact-checking, there has been very limited work on applying multi-modal techniques to fact verification. In this work, we propose a novel unimodal feature-enhanced and cross-modal correlation learning approach (UFCC) for multi-modal fact verification by jointly modeling the basic intra-modal semantic correlation and the inter-modal correlation. Specifically, UFCC consists of a text-semantic feature Module, an image-semantic feature module and a text-image correlation module. In the text-semantic feature module, UFCC exploits pre-trained backbones to separately extract text features from claims and documents. Then we utilize the signed attention mechanism to enhance text information representation by using different text features as query. The image-semantic feature module is similar to the text. In the text-image correlation module, UFCC first adopts the fine-tuned clip model to encode the claims' (or documents') textual and visual features. Then, UFCC explore the cross-modal relationships between the extracted features by using similarity layer. Based on this, we finally fuse the text and image features for better performance. Our team, Zhang, won the fourth prize (F1-score: 77.423%) in Factify challenge hosted by De-Factify2 @ AAAI 2023, which demonstrated the effectiveness of the method.

Keywords

Multi-modal fact verification, attention, fine-tuned clip, De-Factify2

1. Introduction

The emergence of social media has revolutionized the traditional way that people access information online [1]. People enjoy the convenience and efficiency of online social media in sharing information and exchanging ideas [2]. However, spreading disinformation and misinformation in the modern media ecosystem has also become much easier. Many reports pointed out that fabricated stories possibly caused citizens' misconceptions about political candidates [3], manipulated stock prices [4], and threatened public health [5]. Therefore, it is desirable to detect and regulate 'fake news' to promote truthful information on social media platforms.

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

zfr888@mail.ustc.edu.cn (F. Zhang); zq126@mail.ustc.edu.cn (Q. Zhang); jwliu6@ustc.edu.cn (J. Liu); esthersyy.sun@mail.utoronto.ca (E. Sun)

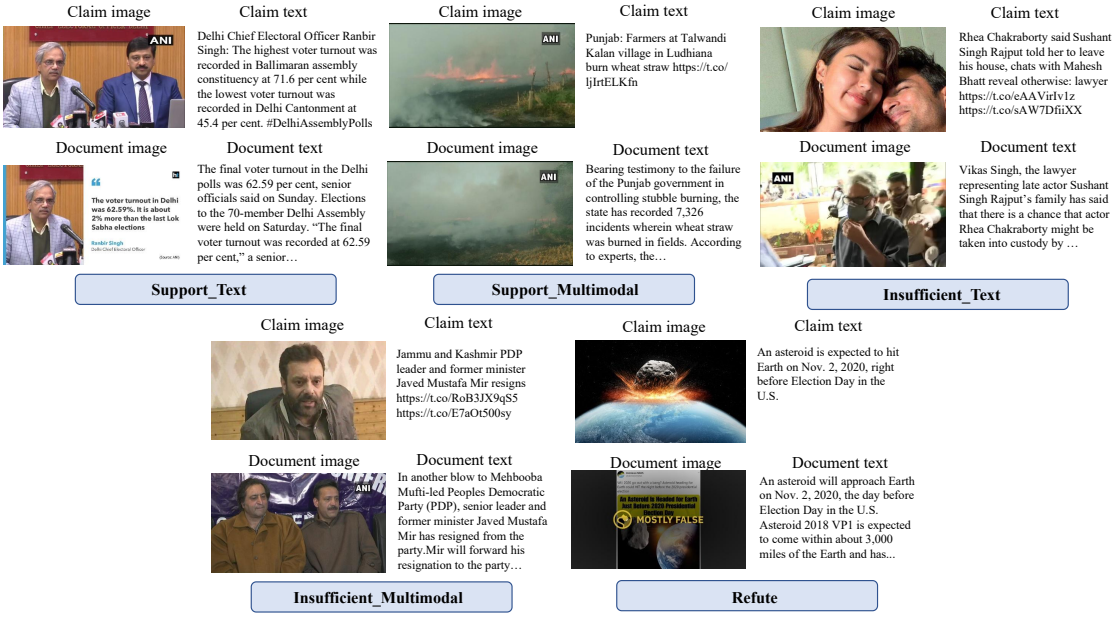

Figure 1: The samples of five categories from the Factify dataset.

Some research formulates this problem as the fact verification task, which targets to automatically verify the integrity of statements using trustworthy corpora [6], e.g., Wikipedia [7]. To verify the claim from the real world in Figure 1, this task first requires searching and retrieving relevant evidence from reliable sources, and making a comparison by absorbing the knowledge of the evidence [8], which would take a great deal of time and effort. Given the large number of claims that need to be checked, manual validation is insufficient and unrealistic. Thus, how to automatically check the integrity of claims, prevent the spread of fake news, and avoid the negative impact on society, is urgently needed for our society. Existing work has presented a number of approaches for tackling fact verification automatically, of which traditional fact verification approaches typically concentrate on text-only content analysis and have produced a range of related work. Nie et al [9] string all the evidence together to verify the claim. Another approach is to reason about each claim evidence pair and aggregate them thus verifying the claim [10, 11]. There are also many fact-checking models that utilize Natural Language Inference(NLI) techniques [12, 13, 14], and one of the most widely used NLI-based models is the Enhanced Sequential Inference Model (ESIM) [12], which uses some form of hard or soft alignment to associate relevant subcomponents between premises and assumptions.

Recently, multi-modal fact verification has received considerable attention since the content form of posts tends to coexist with multi-modal information. Compared to text-based fact verification, multi-modal fact verification is a new and challenging area of research [6]. Although both Image and text contain rich information, they are in heterogeneous modalities and there is a modality gap between them [15]. Some early detectors proposed to investigate cross-modal information, but the lack of the construction of large-scale annotated datasets hindered the

development of the methods [6]. Considering this, the multi-modal fact verification (Factify) challenge at AAAI 2022 has been held [16] and released a large-scale dataset. A series of excellent works have emerged that facilitate the development of multi-modal fact verification systems [17, 18]. For example, Team UofA-Truth [19] directly employs cosine similarity to connect textual (visual) representations of statements and documents. Team Yao [20] uses multiple pre-trained models to extract image and text features, and uses the co-attention mechanism to enhance the information representation. Although the methods mentioned above make use of multi-modal information, they don't effectively enhance the unimodal information nor extract the relationship of inter-modal information

Based on the large-scale and high-quality Defactify dataset [21, 22] and previous excellent works [20, 23, 24], we propose a novel unimodal feature-enhanced and cross-modal correlation learning approach(UFCC) for multi-modal fact verification. Specifically, UFCC consists of a text-semantic feature module, an image-semantic feature module, a text-image correlation module, and a backbone network. In the text-semantic feature module and image-semantic feature module, we first extract claims and documents' unimodality features(text or image) from the pre-trained backbone. These features are further fused by utilizing the signed attention mechanism [25] and a transformer encoder [26] to capture both positive (consistency) and negative (inconsistency) correlations. In the text-image correlation module, we introduce the fine-tuned clip [27] model to reduce the modality gap and reconstruct the claims' (or documents') image and text representations. Finally, we calculate their similarity [28] and map them to the reconstruction features. Figure 2 depicts our framework, showing a 'Support Text' example from the dataset [21] along with the corresponding evidence. As the task requires machine comprehension and modality understanding, we perform different evaluations to design the network and the representations.

The main contributions of this paper are as follows: (1) we propose a novel unimodal feature-enhanced and cross-modal correlation learning approach(UFCC) for multi-modal fact verification by jointly modelling the basic intra-modal semantic correlation and the inter-modal correlation. We effectively enhance the unimodal information and extract the relationship of inter-modal information. (2) We propose two specified semantic feature modules with a signed attention mechanism [25] and a fusion layer to capture both consistency and inconsistency unimodal enhanced features. (3) We introduce a fine-tuned clip model and fusion layer to reduce the claim's modality gap by capturing the relationship between text and image. (4) Our UFCC model outperforms the baseline by at least 12.4% and won the fourth prize in the Factify challenge hosted by De-Factify2 @ AAAI 2023 [21].

2. Related Works

2.1. Fact verification on Text

Fact verification on Text can be seen as a Recognizing Textual Entailment(RTE) task [29, 30], where the goal is to predict whether the text supports or disproves the claim. It is divided into the following main steps, first retrieving documents from a text source(e.g., Wikipedia) that are relevant to a given claim, then selecting sentences that may contain evidence, and finally assigning an authenticity relationship label by the model to support or disprove the claim.

Typical retrieval strategies in the evidence retrieval process include commercial search APIs, Lucene indexes, entity linking, or ranking functions (e.g., dot product of TF-IDF vectors [31]). To improve accuracy, retrieved evidence can also be re-ranked using the stance detection systems acting as a fine-grained filter [32].

Regarding claim verification, much of the recent work is based on graph neural networks. GEAR [33] and KGAT [34] construct graphs with evidence as nodes and use deep graph neural networks for knowledge propagation; DREAM [35] further employ XLNet [36] and build semantic hierarchical graphs for inference to improve performance. These graph-based models establish node interactions for joint inference over several evidence fragments. In recent years, the use of transformer-based linguistic representation models(LRMs) such as BERT [37], RoberTa [38], etc. has also demonstrated their robust performance in claim verification tasks. Transformer-XH [39] propagates knowledge between [CLS] tokens of different evidence fragments; CorefBERT [40] trains a BERT-based LRM which uses an additional target modelling co-reference knowledge and uses it in the KGAT architecture.

2.2. Muti-Modal Fact Verification

Though the majority of existing works have focused on text, some early efforts also investigated how to incorporate multimodal information, Zlatkova [41] created a new dataset and explored some baselines; Lee et al [23] proposed a unifying textual and visual matching layer to confuse the two modality information. They pioneered this new direction.

However, the mentioned methods [41, 23] above all have the problem that the constructed dataset is not large enough for fact verification and the methods cannot effectively learn from the shared information between modalities. Considering this, the multi-modal fact verification (Factify) challenge at AAAI 2022 [42] has been held and a series of excellent works have emerged which facilitate the development of multimodal fact-checking systems. Team Yao [20] adopted the ensemble method by using different pre-trained models and several co-attention modules. Team UofA-Truth [19], used a straightforward approach that concatenated the claim, document textual(visual) representations and their cosine similarity. Team Yet [17] integrated disturbance on the embedding layer, a new loss function, and data augmentation by sequential dropout layers into the vanilla RoBERTa. Gao et al [18] proposed an ensemble model architecture by extracting various information for each modality individually. They applied multiple attention mechanisms to learn the multimodal interaction between visual and textual content pairs.

2.3. Large-scale pre-trained model

Large-scale pre-trained models (PTMs) based on transformer architectures such as BERT [37], ViT [43] and CLIP [27] have demonstrated their powerful performance in a wide range of fields such as NLP, CV, and multimodality. Due to the complex pre-training objectives and large model parameters, PTMs can efficiently capture knowledge from large amounts of labelled and unlabelled data and store the knowledge in a large number of parameters. At the same time, PTMs can be fine-tuned for a specific task and thus applied to a variety of downstream tasks.

In the field of natural language processing, since 2018 we have seen the rise of a range of large-scale pre-trained language models (PLMs) , such as BERT [37], RoBERTa [38], DeBERTa

[44] and others. These PLMs have been fine-tuned by using task-specific labels and have created new levels of skill in many downstream tasks. One of them, Bert, is based on the Multi-Layer Transformer Encoder architecture and uses both Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) methods to train models unsupervised. RoBERTa uses a random mask mechanism and removes the NSP task to enhance Bert's performance. DeBERTa greatly improves the efficiency of pre-training and the performance of downstream tasks by decoupling the attention mechanism, enhanced mask decoder and virtual adversarial training (SiFT).

In the field of computer vision, the ViT [43] model proposed by the Google team applied transformer to image classification, pioneering the application of transformer in vision. The model divides the input image into multiple patches (16x16), then projects each patch into a fixed-length vector to feed the Transformer and uses the final output vector for the classification task. After ViT was proposed, vision models such as DeiT [45], Swin Transformer [46] and others were proposed, which significantly advanced the progress of the computer vision field.

In the field of the multimodal domain, pre-trained models have also demonstrated their powerful performance, such as CLIP [27], ViLT [47], DALL-E 2 [48] and so on. Among them, CLIP is a multimodal pre-training model based on contrasting text-image pairs. The model is trained by jointly training an image encoder and a text encoder to predict the correct pairing of a batch of (image, text) training samples to obtain a transferable visual model, and it has a powerful zero-shot capability that can be used in a variety of downstream tasks.

These pre-trained models have achieved excellent performance in a variety of downstream tasks due to their powerful information characterization capabilities. At the same time, many previous works have applied them to our task and achieved better results. This definitely motivates us to use these pre-trained models as our backbone network to characterize image and text information and convert them into contextual embedding.

3. Method

Muti-modal fact verification in this task is formulated as a five-classification problem to judge the given claims with tests and images entailed in the given documents [20]. To address the problem, we propose a novel unimodal feature-enhanced and cross-modal correlation learning approach(UFCC) by jointly modelling the unimodal semantic correlation and the cross-modal consistency [49]. As illustrated in Figure 2, it consists of a text-semantic correlation Module, an image-semantic correlation module, a text-image consistency Module, and a backbone network. Firstly, given a claim with the text T_C and the image I_C, we denote the document text T_D and the document image I_D. And the task label is defined as y. The goal is to find out the relationship between given claims and documents as follows [22, 21]:

- support_Multimodal: the text and image of the given claim entailed similar news.
- Support_Text: the text of the given claim is entailed similar news but the image is not.
- Insufficient_Multimodal: the image is entailed similar news but the text is not.
- Insufficient_Text: the text of the given claim is not entailed but may have common words and the image is not entailed.

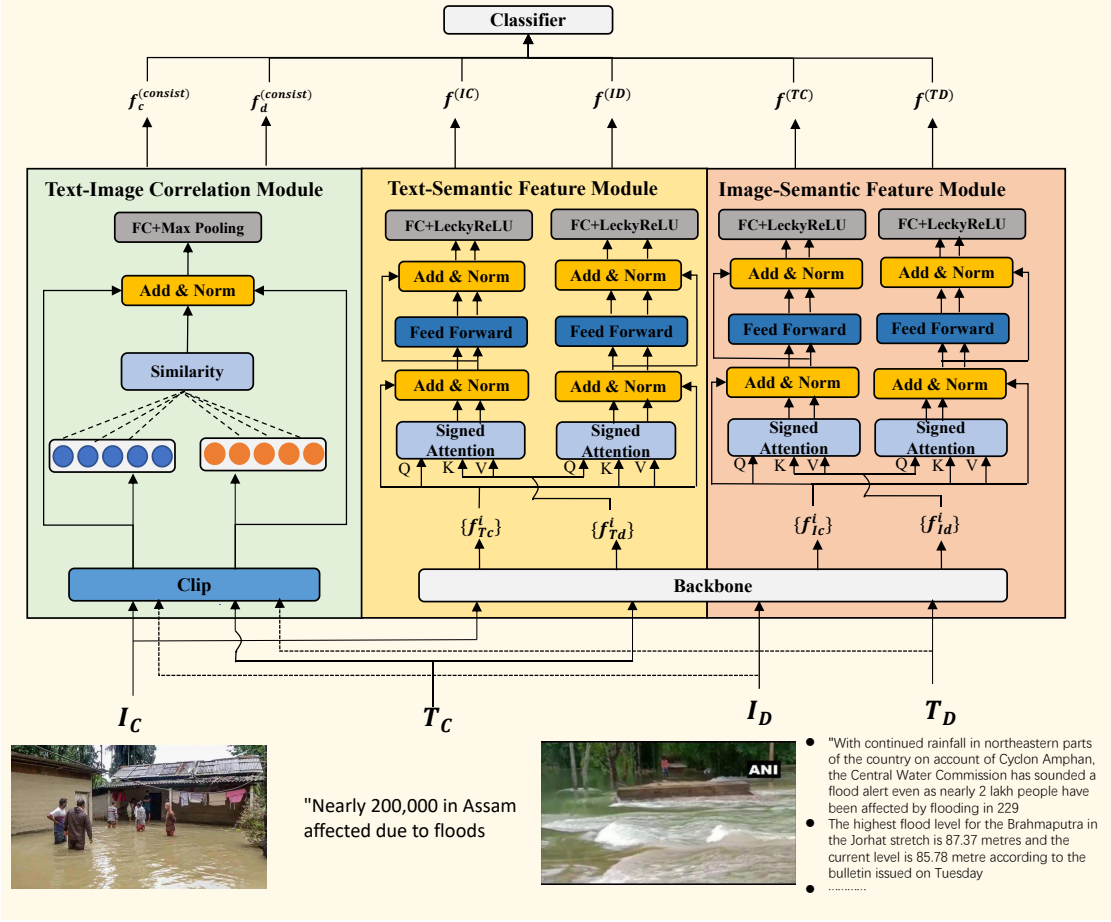

Figure 2: The overall architecture of the proposed UFCC. It contains three components: two unimodal correlation modules, and a text-image consistency correlation module. The cross-modal semantic consistency features and the feature-enhanced unimodal representations are finally concatenated to form the discriminative fact representation.

- Refute: the text and image of the given claim are fake.

3.1. The Backbone Network

The backbone network is designed in a two-stream fashion, consisting of a textual encoder and a visual encoder to extract basic features from the text.

Textual Feature Extractor. Specifically, according to several experiments, we employ a pre-trained DeBERTa model to map the word sequence of T_c into an embedding sequence $\{e_{Tc}^i\}_{i=1}^L$ of length L. The embedding sequence is then further consumed by a bidirectional long short-term memory network (Bi-LSTM) to obtain the final textual feature sequence $\{f_{Tc}^i\}_{i=1}^L \in \mathbb{R}^d$.

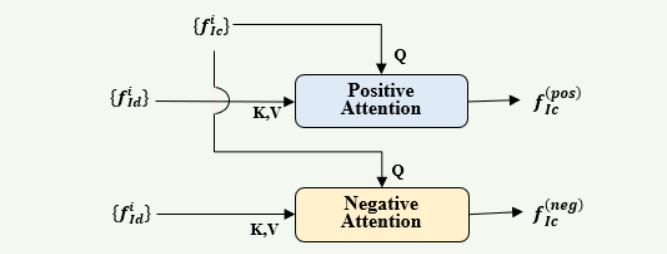

Figure 3: The architecture of the signed attention mechanism

It is formulated as follows:

$$\{f^i_{Tc}\} = W_t^T(\text{Bi-LSTM}(e^i_{Tc})) + b_t \tag{1}$$

where W_t and b_t denote the learnable parameters of a fully connected (FC) layer.

Visual Feature Extractor. Based on the excellent performance of the Vit [43] model, we adopt a pre-trained Vit model [43] to encode image I_c into an initial feature map with dimension 768, which is then transformed by a linear block into the visual feature sequence $\{f^i_{Ic}\}^L_{i=1} \in \mathbb{R}^d$. This is formulated as follows:

$$\{f^i_{Ic}\} = \text{MLP}(\text{ViT}(I_c)) \tag{2}$$

where ViT denotes the pre-trained Vit model, MLP represents multi-Layer perceptrons. Similarly, we get the representation of the document text and document image $\{f^i_{Td}\}$,$\{f^i_{Id}\}$.

3.2. Unimodal Semantic Feature Module

Our UFCC model has two unimodal semantic feature modules including the text-semantic feature module and the image-semantic feature module. Taking the text modules as an example, we first further fused two extracted features f_{Tc},f_{Td} by utilizing the signed attention mechanism [25], in order to simultaneously capture both semantic consistency and inconsistency correlations [49, 50]. In the traditional attention mechanism, if the correlations between query and keys are negative, we would treat it as insignificant [26]. However, such a negative correlation may represent the inconsistency semantics that can be beneficial to the fact verification task [49]. As shown in Figure 3, it takes the f_{Tc} as the query, and f_{Td} as the key and value to calculate the positive correlation as follows:

$$\begin{aligned} \alpha^i_{pos} &= \text{Softmax}\left((f^i_{Tc})^T(f^i_{Td})/\sqrt{2d_e}\right) \\ f^{pos}_{Tc} &= \left(\sum_{i=1}^{L}\alpha^i_{pos}f^i_{Td}\right)\Big/\left(\sum_{i=1}^{L}\alpha^i_{pos}\right) \end{aligned} \tag{3}$$

where $2d_e$ is the dimension of $f^i_{T_c}$. α^i_{pos} denotes the positive attention coefficients. A larger α^i_{pos} indicates that the query claim text is more positively semantically associated with the document text.

On the contrary, the negative attention mechanism utilizes a “-softmax" operation to amplify the inconsistent correlations [49, 50, 25, 26], that is,

$$\begin{aligned} \alpha_{neg}^{i} &= -\text{Softmax}\left(-(f_{Tc}^{i})^{T}(f_{Td}^{i})/\sqrt{2d_e}\right) \\ f_{Tc}^{neg} &= \left(\sum_{i=1}^{L}\alpha_{neg}^{i}f_{Td}^{i}\right)\Big/\left(\sum_{i=1}^{L}\alpha_{neg}^{i}\right) \end{aligned} \tag{4}$$

Then we utilize two residual modules and a two layers feedforward Neural Network to enhance the features:

$$\begin{aligned} f_{Tc}^{p} &= f_{Tc}^{p} + \text{FFN}(f_{Tc}^{p}) \\ f_{Tc}^{n} &= f_{Tc}^{n} + \text{FFN}(f_{Tc}^{n}) \end{aligned} \tag{5}$$

where FFN denotes a two layers feedforward Neural Network. We then fuse the updated text representations with two f_{Tc}^{p},f_{Tc}^{n} arithmetic operations, and obtain the final representation of the text evidence $\{f_{Td}^{i}\}$ towards the text claim $\{f_{Tc}^{i}\}$ based on a linear layer.

$$f^{(TC)} = \sigma\left(W_a\left[f_{Tc}^{p} : f_{Tc}^{n}\right] + b_a\right) \tag{6}$$

where [:] denotes concatenation operation. W_a and b_a are learnable parameters for aggregating the representations.σ denotes the LeckyReLU activation function.

Similar to Eqn. 3-6, we get the outputs of the claim image $f^{(IC)}$; the document text $f^{(TD)}$, and the document image $f^{(ID)}$.

3.3. Cross-modal Correlation Module

The text and image features extracted by DeBERTa and ViT respectively have significant semantic modal gaps, and it is difficult for the network to learn their consistency correlation if they are fused directly. Therefore, we extract the alignment features of the text-image pair by utilizing and fine-tuning the CLIP model. The extracted feature is denoted as $\{f_{clip-T}^{i}\}$ and $\{f_{clip-I}^{i}\}$.

During fine-tuning, we pass the text-image pair through the CLIP encoders and normalize their embeddings. We produce a joint embedding that is a dot product of the image and text ones. Merely combining the CLIP-based features with the multimodal features cannot necessarily provide enough reliable information. The reason is that fact verification is not completely correlated with image-text correlation [24, 23]. To address the ambiguity issue between multimodal features, we measure the cosine similarity between the text features and the image features provided by CLIP, to adjust the intensity of fused features [24]. The cosine similarity and the fused module are calculated as follows.

$$\text{sim} = \text{Norm}\left(\frac{f_{clip-T}\cdot\left(f_{clip-I}\right)^{T}}{\|f_{clip-T}\|\,\|f_{clip-I}\|}\right). \tag{7}$$

$$f_{clip} = \text{sim}\cdot\left(f_{clip-T}\cdot f_{clip-I}\right) \tag{8}$$

We then concatenate all previous features into one feature and update the final claim inter-modal representations based on MaxPooling.

$$\begin{aligned}\tilde{f}_c^{(consist)} &= \sigma\left(W_c\left[f_{clip} : f_{clip-T} : f_{clip-I}\right] + b_c\right) \\ f_c^{(consist)} &= \text{Max Pooling}\left(\tilde{f}_c^{(consist)}\right)\end{aligned} \tag{9}$$

where [:] denotes concatenation operation. W_c and b_c are learnable parameters for aggregating the representations.σ denotes LeckyReLU activation function. Similarly, we also get the final document inter-modal representations $f_d^{(consist)}$.

3.4. Classifier

UFCC leverages the unimodal semantic correlation as well as the multimodal consistency correlation to jointly perform fact classification. The cross-modal semantic consistency features $f_d^{(consist)}$ and $f_c^{(consist)}$ and the feature-enhanced unimodal representations $f^{(IC)}$, $f^{(ID)}$, $f^{(TC)}$ and $f^{(ID)}$ are finally concatenated to form the discriminative fact representation, which is further transformed by an FC layer with softmax activation function to verify the fact as the following:

$$\hat{y} = \sigma_1(W_d^T[f_d^{(consist)} : f_c^{(consist)} : f^{(IC)} : f^{(ID)} : f^{(TC)} : f^{(ID)}] + b_d) \tag{10}$$

where W_d and b_d are the parameters of the classifier layer. σ_1 denotes the Sotfmax activation function. The overall framework is supervised with a cross-entropy criterion.

$$\mathrm{L} = -\sum_{i=1}^{|M|} y_i \log\left(\hat{y}_i\right) \tag{11}$$

4. Experiments

4.1. Experimental Settings

Dataset. Factify is a dataset for multi-modal fact verification whose goal is to identify if the claim entails the document. In this dataset, each sample includes claim image, claim, claim ocr, document image, document, document ocr, and category. Also, the dataset contains a total of 5 categories, which contain Support_Text, Support_Multimodal, Insufficient_Text, Insufficient_Multimodal and Refute. The dataset is divided into three parts: a training set, a validation set, and a test set. The training set contains a total of 35,000 samples, of which 7,500 are for each category. The validation set contains 7,500 samples, of which 1,500 are for each category. The test set is used to evaluate the performance of the model and contains a total of 7,500 samples. For more details, we refer readers to [**?**].

Implementation Details. In the textual encoder, we set the length of the input text to at most 128 words, and utilize the pre-trained DeBERTa [44] model to initialize the word embeddings with 768 dimensions. In the visual encoder, we use the pre-trained Vit model [43] to extract the visual feature. The dimension of the visual feature $\{f_{Ic}^i\}$ and textual feature $\{f_{Tc}^i\}$ are 256. In terms of parameter setting, we set the learning rate of the overall framework to $1.8e^{-4}$, and fine-tune the DeBERTa model and the Vit model with a learning rate of $5e^{-5}$. The batch size of

Rank	Team	Support _ Text (%)	Support _ Multimodal (%)	Insufficient _Text (%)	Insufficient _Multimodal (%)	Refute (%)	Final (%)
1	Triple-Check	82.767	91.383	85.19	89.217	1	81.82
2	INO	81.235	90.029	88.807	85.233	99.933	80.795
3	Logically	80.383	90.511	84.393	85.627	98.512	78.967
4	**Zhang**	**76.645**	**87.851**	**81.61**	**87.934**	**99.933**	**77.423**
5	gzw	78.493	86.321	81.423	83.269	1	76.051
6	coco	77.25	86.493	81.517	82.995	1	75.697
7	Noir	77.1	87.26	78.493	81.563	99.699	74.522
8	Yet	70.745	82.629	78.592	71.904	1	69.085
-	Baseline	50	82.721	80.24	75.931	98.82	64.99

Table 1
Factify Official Leaderboard (Our team, Zhang, won the fourth prize in Factify challenge)

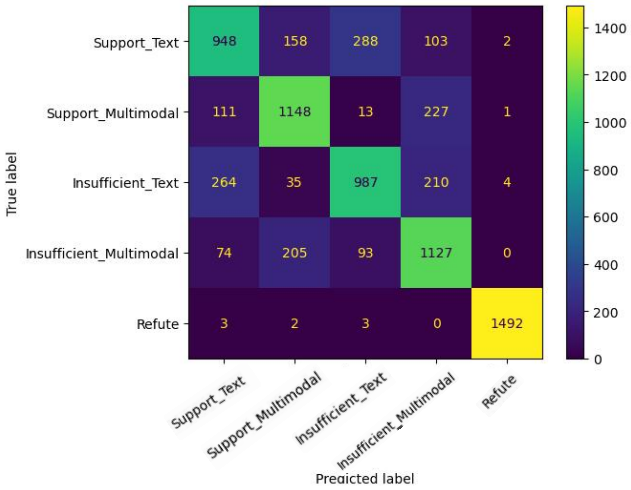

Figure 4: Confusion matrix of the testing set.

the input is 16. Adam optimizer is used to train UFCC. Weighted F1 scores are employed as the evaluation metrics.

4.2. Analysis of experimental results

Testing Performance. The final results of the test are presented in Table 1. the average F1 score value of our proposed model was 77.423%, outperforming the baseline by 12.4%, while achieving fourth place among all participating teams. We can attribute the strengths of UFCC to the following aspects: (1) the use of Vit and DeBERTa as the backbone network, resulting in an excellent representation of image and text information. (2) The use of the Signed Attention mechanism captures both semantic consistency and inconsistent correlation, which in turn enhances unimodal information. (3) Using CLIP to fuse multimodal information and adjust the strength of the fusion by calculating the cosine similarity of text and image information.

Model	-w/o TIC	-w/o TSF	-w/o ISF	-w/o Sim	-w/o S-A	UFCC
F1 (%)	73.17	69.12	74.45	76.49	75.09	**77.42** (+0.41)

Table 2
Ablation study of our model in terms of test score.

Confusion Matrix. Figure 4 shows the confusion matrix of the testing set. We can find that the model can accurately discriminate the Refute category, but performs poorly in the Support_Text and Insufficient_Text categories, indicating that the model is weak at discriminating when the claimed image or text is neither supported nor refuted. In subsequent studies, we will improve it for this purpose.

4.3. Ablation Study

To show the effectiveness of different modules in UFCC, we compare it with the sub-models "-w/o TIC", "-w/o TSF", "-w/o ISF" and "-w/o Sim". They denote the variant of UFCC without considering the text-image correlation module, the text-semantic feature module, the image-semantic feature module and the clip similarity module, respectively. And the "-w/o S-A" denotes UFCC without replacing the signed attention to the traditional attention mechanism. The comparison results are shown in Table 2. We can observe that all ablation variants perform worse than the complete UFCC model on the Factify2 dataset. The results indicate that: (1) All three modules are important for fact verification; (2) the modal similarity architecture can facilitate the multi-modal fusion; (3) The signed attention mechanism can both capture unimodal consistency and inconsistency correlations.

5. Conclusion

In this paper, we propose a novel unimodal feature-enhanced and cross-modal correlation learning approach(UFCC) for multi-modal fact verification by jointly modelling the basic intra-modal semantic correlation and the inter-modal. Our framework enables more effective multimodal fusion by introducing fine-tuned clip and similarity-fused layers. Besides, we adopt the signed attention mechanism to enhance the unimodal information representation. At last, the ablation study demonstrates the effectiveness of our proposed approach.

In " Hand movement appears to help in teaching about statistical models", Matthew Hutson explains that gesture movements help speakers convey ideas, think and learn.

First, the authors pointed out that psychological research is exploring the possibility of gesturing while learning. Studies have shown that learners who imitate a teacher's movements are better at helping them learn and remember, even if they don't know why.

And the authors mention a study that tests this hypothesis. New work extends this finding. The researchers tested the subconscious effects of

References

[1] 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 Conference on Computer Vision and Pattern Recognition, 2022, pp. 14940–14949.

[2] R. Mishra, Fake news detection using higher-order user to user mutual-attention progression in propagation paths, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 652–653.

[3] H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election, Journal of economic perspectives 31 (2017) 211–36.

[4] S. Kogan, T. J. Moskowitz, M. Niessner, Fake news: Evidence from financial markets, Available at SSRN 3237763 (2019).

[5] Ü. Recep, A. Ş. ÇİÇEKLİOĞLU, Fake news pandemic: Fake news and false information about covid-19 and an analysis on factchecking from turkey in sample teyit. org, Erciyes İletişim Dergisi 9 (2022) 117–143.

[6] Z. Guo, M. Schlichtkrull, A. Vlachos, A survey on automated fact-checking, Transactions of the Association for Computational Linguistics 10 (2022) 178–206.

[7] D. Milne, I. H. Witten, Learning to link with wikipedia, in: Proceedings of the 17th ACM conference on Information and knowledge management, 2008, pp. 509–518.

[8] B. Adair, C. Li, J. Yang, C. Yu, Progress toward "the holy grail": The continued quest to automate fact-checking, in: Computation+ Journalism Symposium,(September), 2017.

[9] Y. Nie, H. Chen, M. Bansal, Combining fact extraction and verification with neural semantic matching networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 6859–6866.

[10] J. Luken, N. Jiang, M.-C. de Marneffe, QED: A fact verification system for the FEVER shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 156–160. URL: https://aclanthology.org/W18-5526. doi:10.18653/v1/W18-5526.

[11] T. Yoneda, J. Mitchell, J. Welbl, P. Stenetorp, S. Riedel, UCL machine reading group: Four factor framework for fact finding (HexaF), in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 97–102. URL: https://aclanthology.org/W18-5515. doi:10.18653/v1/W18-5515.

[12] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, D. Inkpen, Enhanced lstm for natural language inference, arXiv preprint arXiv:1609.06038 (2016).

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. URL: https://aclanthology.org/N18-1202. doi:10.18653/v1/N18-1202.

[15] B. M. Yao, A. Shah, L. Sun, J.-H. Cho, L. Huang, End-to-end multimodal fact-checking and

explanation generation: A challenging dataset and models, arXiv preprint arXiv:2205.12487 (2022).
[16] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. N. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, Factify: A multi-modal fact verification dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, ceur, 2022.
[17] Y. Zhuang, Y. Zhang, Yet at factify 2022 : Unimodal and bimodal roberta-based models for fact checking (short paper), in: DE-FACTIFY@AAAI, 2022.
[18] J. Gao, H.-F. Hoffmann, S. Oikonomou, D. Kiskovski, A. Bandhakavi, Logically at factify 2022: Multimodal fact verfication, ArXiv abs/2112.09253 (2021).
[19] A. Dhankar, O. R. Zaiane, F. Bolduc, Uofa-truth at factify 2022 : A simple approach to multi-modal fact-checking, in: DE-FACTIFY@AAAI, 2022.
[20] W.-Y. Wang, W.-C. Peng, Team yao at factify 2022: Utilizing pre-trained models and co-attention networks for multi-modal fact verification (short paper), ArXiv abs/2201.11664 (2022).
[21] S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A. Reganti, 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, 2022.
[22] 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, 2022.
[23] N. Vo, K. Lee, Where are the facts? searching for fact-checked information to alleviate the spread of fake news, arXiv preprint arXiv:2010.03159 (2020).
[24] S. K. Kiran, M. Shashi, K. Madhuri, Multi-stage transfer learning for fake news detection using awd-lstm network (2022).
[25] J. Huang, H. Shen, L. Hou, X. Cheng, Signed graph attention networks, in: International Conference on Artificial Neural Networks, Springer, 2019, pp. 566–577.
[26] 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).
[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763.
[28] A. Tversky, Features of similarity., Psychological review 84 (1977) 327.
[29] I. Dagan, B. Dolan, B. Magnini, D. Roth, Recognizing textual entailment: Rational, evaluation and approaches–erratum, Natural Language Engineering 16 (2010) 105–105.
[30] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, arXiv preprint arXiv:1508.05326 (2015).
[31] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, A. Mittal, The fact extraction and VERification (FEVER) shared task, in: Proceedings of the First Workshop on Fact Extraction and VERification (FEVER), Association for Computational

Linguistics, Brussels, Belgium, 2018, pp. 1–9. URL: https://aclanthology.org/W18-5501. doi:10.18653/v1/W18-5501.
[32] 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).
[33] J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Gear: Graph-based evidence aggregating and reasoning for fact verification, arXiv preprint arXiv:1908.01843 (2019).
[34] Z. Liu, C. Xiong, M. Sun, Kernel graph attention network for fact verification (2019).
[35] W. Zhong, J. Xu, D. Tang, Z. Xu, N. Duan, M. Zhou, J. Wang, J. Yin, Reasoning over semantic-level graph for fact checking, arXiv preprint arXiv:1909.03745 (2019).
[36] 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).
[37] 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).
[38] 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).
[39] C. Zhao, C. Xiong, C. Rosset, X. Song, P. Bennett, S. Tiwary, Transformer-xh: Multi-evidence reasoning with extra hop attention (2020).
[40] D. Ye, Y. Lin, J. Du, Z. Liu, P. Li, M. Sun, Z. Liu, Coreferential reasoning learning for language representation, arXiv preprint arXiv:2004.06870 (2020).
[41] D. Zlatkova, P. Nakov, I. Koychev, Fact-checking meets fauxtography: Verifying claims about images, arXiv preprint arXiv:1908.11722 (2019).
[42] 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.
[43] 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).
[44] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).
[45] 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.
[46] 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.
[47] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.
[48] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, M. Chen, Hierarchical text-conditional image generation with clip latents, arXiv preprint arXiv:2204.06125 (2022).
[49] M. Sun, X. Zhang, J. Ma, Y. Liu, Inconsistency matters: A knowledge-guided dual-

inconsistency network for multi-modal rumor detection, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 1412–1423.

[50] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, Y. Zhang, Mfan: Multi-modal feature-enhanced attention networks for rumor detection (2022).