LyRE:Learning Varying Fusion Degrees with Hierarchical Aggregation to Improve Multimodal Misinformation Detection

Yidu Chen[1,2,3], Bo Ma[1,2,3(✉)], Yating Yang[1,2,3(✉)], Dilxat Abdureyim[1,2,3], Rui Dong[1,2,3], Zhen Wang[1,2,3], Lei Wang[1,2,3], and Zhou Xi[1,2,3]

[1] Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi, China
{mabo,yangyt,dilxat,dongrui,wang_zhen,wanglei,zhouxi}@ms.xjb.ac.cn
[2] University of Chinese Academy of Sciences, Beijing, China
chenyidu22@mails.ucas.ac.cn
[3] Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi, China

Abstract. The rapid proliferation of misinformation poses serious concerns, necessitating the development of efficient and accurate automated detection methods. Existing multimodal misinformation detection approaches predominantly focus on fusing information from different modalities. However, the diverse nature of multimodal posts on social media means that solely focusing on fusion can introduce noise, particularly in posts with weak inter-modal correlations. To address this challenge and effectively handle diverse misinformation instances, we propose a novel method ***L****earning Vary****y****ing Fusion Degrees with Hierarchical Aggr****e****gation*(LyRE). LyRE employs classifiers at different stages of a hierarchical fusion process, enabling the model to learn from representations with varying degrees of cross-modal interaction and adapt to different types of multimodal data. Experimental results on multiple publicly misinformation detection datasets demonstrate that LyRE outperforms other state-of-the-art and highly competitive misinformation detection methods.

Keywords: Misinformation Detection · Multimodal Misinformation Detection · multimodal fusion · Hierarchical Aggregation

1 Introduction

Misinformation mimics the form and style of legitimate news media but fabricates information to mislead the public for malicious purposes[15], raising serious concerns over its rapid proliferation on social media platforms[17]

The evolution of social media has significantly reduced the cost of information dissemination, driving a shift from text-based news posts to multimodal formats[14]. This shift is particularly evident in the proliferation of multimodal misinformation, where eye-catching images are often coupled with misleading

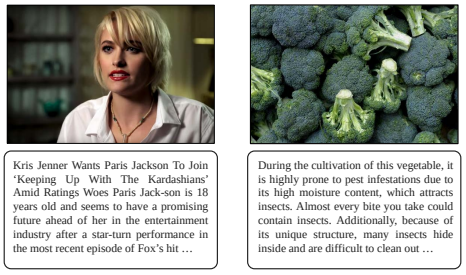
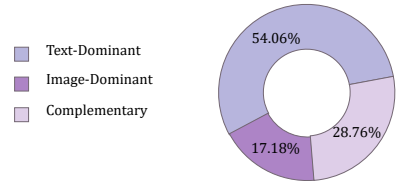
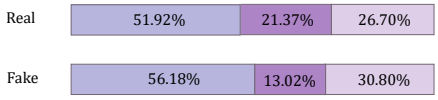


Fig. 1: Examples of misinformation on social media. The left instance where image and text present relatively isolated information, allowing for classification based solely on the text. The right instance where the textual information is incomplete, the text alone provides insufficient information to discern the specific "*vegetable*" being referenced, necessitating image analysis for comprehensive understanding



(a) The overall proportion of all classes.



(b) Respective Proportions of Real and Fake Classes

Fig. 2: Analysis of the Proportions of Text-Dominant, Image-Dominant, and Complementary Instances within the Weibo-21 Dataset.

text to create engaging content that users are more likely to share [1]. Fabricated for economic or political gain, misinformation often relies on text to convey its primary claims, while images and videos provide supporting visual cues [23]. Automated multimodal misinformation detection systems leverage this interplay by analyzing semantic, emotional, and textural features extracted from both text and visual modalities to discern factual content from falsehoods.

Early fusion of modality features is a prominent approach in multimodal processing. Existing early fusion methods for multimodal misinformation detection primarily include feature concatenation [25], cross-attention mechanisms [28,35], semantic alignment-based fusion [33],multi-view fusion[30], and entity relation-based fusion [7]. While these methods effectively integrate features from two modalities to some extent, early fusion at the feature level may rely heavily on highly correlated data, limiting its applicability to all types of misinformation posts. First, misinformation can be unimodal-dominant or multimodal complementary, Figure 1 shows two misinformation samples, and Figure 2 depicts the proportions of modal dominant, suggesting that a large portion exhibits text dominance, while a smaller portion shows image dominance. Many of these instances contain irrelevant content in the non-dominant modality, with a weak inter-modal correlation. Early fusion in such cases might introduce noise, hindering accurate classification. Second, highly fusing visual and textual information can lead to information loss due to inherent semantic differences[24]. Third, Con-

catenation and attention-based fusion methods often result in quadratic time and space complexity [19], requiring significant computational resources. To address the aforementioned challenges, we explore an effective hierarchical fusion approach to enhance multimodal misinformation detection, particularly in scenarios with varying degrees of inter-modal correlation. Inspired by the neural dynamics observed in the human brain during multisensory processing we propose a ***L****earning Var****y****ing Fusion Degrees with Hierarchical Agg****re****gation*(LyRE) multimodal misinformation detection framework.

The neuroscientific studies have revealed a gradual integration process when the brain handles multisensory stimuli, unimodal stimuli appearing first followed by multimodal integration within a short time window [3,12]. This suggests that multisensory fusion in the brain may be a progressive phenomenon, where the brain first processes and responds to unimodal information, and then processes and responds to the integrated multimodal information in a hierarchical manner[21]. Similar to the human brain, the LyRE framework has two core operations: 1) Hierarchical Aggregation and 2) Varying Fusion Degrees Learning. In Hierarchical Aggregation, unimodal features are initially mapped and outputted in a simple manner. Subsequently, these features are progressively fused with increasing degrees of cross-modal features, layer by layer. Although some information loss occurs during feature fusion, we retain the results at each step to mitigate this loss. In Varying Fusion Degrees Learning, we use independent classifiers to learn information from different perspectives of fusion degrees, adapting to various types of multimodal data. The connected network is composed of simple MLP units. Despite the simplicity of LyRE, it achieves remarkable performance.

Our contributions can be summarized as follows:

1. We propose a simple but effective module for cross-modal feature aggregation that achieves efficient cross-modal information fusion using reusable network pathways and weight addition, ensuring computational efficiency.
2. We introduce a method that focuses on information with varying degrees of cross-modal fusion, addressing misinformation posts with different cross-modal relevance.
3. The proposed LyRE method achieves state-of-the-art performance on multimodal fake information detection datasets Gossipcop and Weibo-21. Moreover, we conduct detailed and comprehensive experimental analyses of feature fusion at different fusion degrees.

2 The Approach

Our proposed method aims to utilize a simple fusion architecture to capture more comprehensive information from images and text in misinformation detection, enabling effective real-time identification. As illustrated in Figure 3, images and text are independently represented using pre-trained models. These representations are then processed through a three-layer network structure, where the degree of cross-modal fusion gradually increases at each layer. We integrate

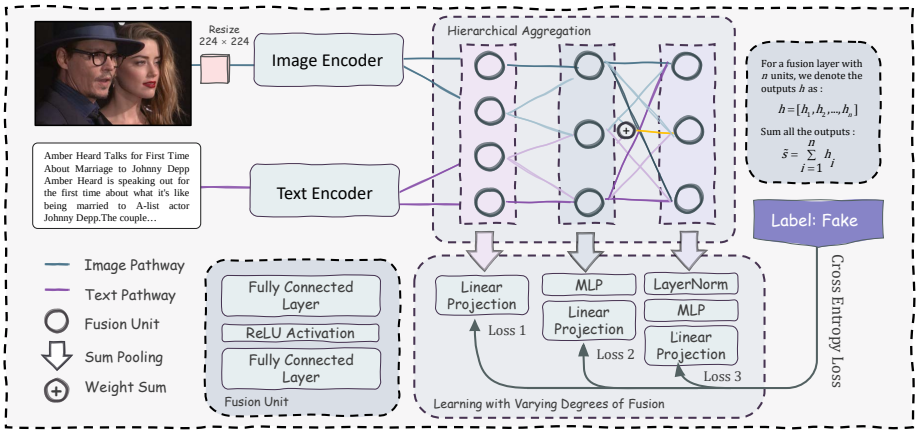


Fig. 3: The framework of LyRE comprises two primary modules: a Hierarchical Aggregation module, consisting of multiple independent Aggregation Units, and a Varying Degrees Classifier module.

vector representations with varying degrees of fusion and design a multi-view classification loss. The details of our proposed method will be introduced in subsequent sections.

2.1 Feature Encoder

Given a multimodal sample $x = \{T, I\}$ as input, we use the pre-trained language model BERT [5] to encode the text. Specifically, given a sentence T, we use a tokenizer of BERT to tokenize T into tokens, and then use max pooling to obtain the text representation, and for the image modality, we utilize the pre-trained vision model MAE [6] as the image encoder. We resize images of different sizes to a fixed size, and then encode the image using MAE :

$$e_t = MaxPooling(BERT(Tokenize(T))) \tag{1}$$

$$e_i = MAE(Resize(I)) \tag{2}$$

where e_t is the text representation, $MaxPooling(\cdot)$ is a max pooling operation, $Resize(\cdot)$ is an operation that resizes the image to 224×224 size, and e_i is an image representation with the same dimension as e_t.

2.2 Hierarchical Aggregation

Neurodynamic studies of the brain have found that as neurons integrate different sensory inputs, the connection strength between neurons in various sensory regions gradually increases during cross-sensory fusion [20]. Inspired by this, when designing the pathways for network connections, the connection strength

between layers in the fusion network gradually increases, leading to a more consistent representation space across modalities. This progressive fusion approach enables a more complete capture of cross-modal information.

Aggregation Unit An aggregation unit consists of two fully connected layers and an activation function:

$$o = \sigma(W_v f(z) + b_v) \tag{3}$$

where $f(\cdot)$ denotes a fully connected layer with one layer, $\sigma(\cdot)$ denotes the ReLU activate function , z denotes the input to the aggregation unit, and W_v, b_v are the parameters of aggregation unit.

Hierarchical Aggregation Layer As presented in Figure 3, we employ a three-level Aggregation Layer to fuse cross-modal information.

In the first level, each aggregation unit is utilized only once, processing a single input and generating a single output, the aggregation units within the second level are each utilized twice, participating in two separate input-output operations. This utilization pattern is further extended in the third level, where each Unit is employed three times. This hierarchical structure with increasing unit utilization facilitates a gradual and increasingly sophisticated fusion of cross-modal information.

Given the image feature e_i and text feature e_t after encoding, the information pathway of the first aggregation layer can be described as follows :

$$o_{1n} = \sigma(f_{1n}(e_i) + b_{v_{1n}}), n \in \{1, 2\} \tag{4}$$

$$o_{1m} = \sigma(f_{1m}(e_t) + b_{v_{1m}}) + b_{v_{1m}}, m \in \{3, 4\} \tag{5}$$

Where $o_{ij}, 1 \leq j \leq 4, 1 \leq i \leq 3$ represent the j aggregation outputs of i layer outputs, f_{kv} denotes the fully connected parameters of k unit in v aggregation layer. Note that e_i and e_t are duplicated to construct outputs under different transformations, ensuring each unit in the subsequent layer receives two distinct inputs.

$$o_{21} = \sigma(f_{21}(o_{11}) + b_{v_{21}}) \tag{6}$$
$$o_{2n} = \sigma(f_{22}(o_{12}) + b_{v_{22}}), \quad n \in \{2, 3\} \tag{7}$$
$$o_{2m} = \sigma(f_{22}(o_{13}) + b_{v_{23}}), \quad m \in \{4, 5\} \tag{8}$$
$$o_{26} = \sigma(f_{23}(o_{14}) + b_{v_{24}}) \tag{9}$$

In the second layer, o_{12} carries information derived from the image, while o_{13} carries information from the text, note that o_{12}, o_{13} are duplicated used.

$$\begin{bmatrix} o_{31} \\ o_{32} \\ o_{33} \\ o_{34} \\ o_{35} \\ o_{36} \\ o_{37} \\ o_{38} \\ o_{39} \end{bmatrix} = \sigma(\begin{bmatrix} f_{31}(o_{21}) \\ f_{31}(o_{23}) \\ f_{31}(o_{26}) \\ f_{32}(o_{22}) \\ f_{32}(o_{23+24}) \\ f_{32}(o_{25}) \\ f_{33}(o_{21}) \\ f_{33}(o_{24}) \\ f_{33}(o_{26}) \end{bmatrix} + \begin{bmatrix} b_{o_{31}} \\ b_{o_{31}} \\ b_{o_{31}} \\ b_{o_{32}} \\ b_{o_{32}} \\ b_{o_{32}} \\ b_{o_{33}} \\ b_{o_{33}} \\ b_{o_{33}} \end{bmatrix}) \tag{10}$$

In the third layer, each neuron participates in the process of modality information aggregation. The outputs o_{21}, o_{22}, o_{23} carry visual information, while o_{24}, o_{25}, o_{26} carry textual information. We performed an orderly combination, ensuring that each neuron passes through both the text and image streams.

2.3 Learning with Varying Degrees of Fusion

Different layers within a neural network learn distinct levels of information and exhibit varying sensitivities to specific features [32]. Lower layers tend to capture more general and abstract features, while higher layers learn more specific and task-oriented representations [31]. Recent studies on multimodal tasks have demonstrated that incorporating modality-specific output heads can enhance the classification performance of fused representations [30]. Inspired by these findings, we introduce independent classifier for representations at different stages of fusion.

The complexity of these classifier increases progressively with the level of fusion. For representations with low fusion, we employ a simple linear transformation. For moderately fused representations, we utilize a multi-layer perceptron (MLP). Finally, for highly fused representations, we incorporate Layer Normalization to mitigate fluctuations arising from the summation of multiple information sources.

Varying Degrees Classifier After obtaining multimodal representations with varying degrees of fusion through summation, we employ three independent classifiers to perform classification on these representations. This approach allows us to capture feature information at different levels of abstraction:

$$r_i = \sum_{j=1}^{2n_i} o_{ij} \quad , i = 1, 2, 3 \tag{11}$$

$$\hat{y_1} = Softmax(W_{p_1} r_1) \tag{12}$$

$$\hat{y_2} = Softmax(MLP(W_{p_2} r_2)) \tag{13}$$

$$\hat{y_3} = Softmax(LayerNorm(MLP(W_{p_3} r_3))) \tag{14}$$

Table 1: Performance comparison on the Gossipcop, Politifact and Weibo-21 datasets. The best performance is highlighted in bold, while underlining highlights the follow-up.

Datasets	Methods	Accuracy	Real News			Fake News		
			Precision	Recall	F1	Precision	Recall	F1
Gossipcop	SAFE [34]	0.838	0.857	0.937	0.895	0.758	0.558	0.643
	SpotFake [25]	0.858	0.866	0.962	0.914	0.732	0.372	0.494
	EANN [27]	0.864	0.887	0.956	0.920	0.702	0.518	0.594
	CAFE [4]	0.867	0.887	0.957	0.921	0.732	0.490	0.587
	BMR [30]	0.890	0.896	**0.977**	0.935	**0.844**	0.521	0.645
	LyRE(Ours)	**0.897**	**0.916**	0.960	**0.937**	0.793	**0.631**	**0.702**
Politifact	SAFE [34]	0.874	0.851	0.830	0.840	0.889	0.903	0.896
	SpotFake [25]	0.760	0.933	0.778	0.848	0.310	0.643	0.419
	CAFE [4]	0.864	0.895	0.919	0.907	0.724	0.778	0.750
	BMR [30]	0.885	0.904	0.917	0.917	0.806	0.781	0.781
	LyRE(Ours)	**0.942**	**0.972**	**0.947**	**0.960**	**0.871**	**0.931**	**0.900**
Weibo-21	SpotFake [25]	0.851	0.786	**0.964**	0.866	**0.953**	0.733	0.828
	EANN [27]	0.870	0.841	0.912	0.875	0.902	0.825	0.862
	CAFE [4]	0.882	0.907	0.844	0.876	0.857	0.915	0.885
	BMR [30]	0.923	0.941	0.905	0.923	0.906	0.942	0.924
	LyRE(Ours)	**0.947**	**0.950**	0.946	**0.948**	0.944	**0.949**	**0.947**

where $\hat{y_1}$, $\hat{y_2}$, and $\hat{y_3}$ represent the predicted probability distributions over the classes for each level of representation fusion.W_{p_1}, W_{p_2}, and W_{p_3} are learnable weight matrices, and MLP denotes a multi-layer perceptron.

We aggregate the predictions from different levels of representation and employ a cross-entropy loss to train the classification objective:

$$\mathcal{L}_i = -\mathcal{E}_{y\sim\hat{Y}_i}[y\log(\hat{y_i}) + (1-y)\log(1-\hat{y_i})], 1 \le i \le 3 \tag{15}$$

$$\mathcal{L}_{total} = \alpha\mathcal{L}_1 + \beta\mathcal{L}_2 + \gamma\mathcal{L}_3 \tag{16}$$

where y denotes the ground-truth label of the input post, and α, β, γ are hyperparameters controlling the weights of the losses from different fusion levels.

During inference, we aggregate the three predicted outputs as the final prediction $\hat{y} = \alpha\hat{y_1} + \beta\hat{y_2} + \gamma\hat{y_3}$.

3 Experiments

3.1 Experimental Setup

We evaluate the performance of LyRE and baselines on three real-world misinformation datasets collected from social media, Gossipcop[22] , Politifact[22], Weibo-21[13]. To assess broader applicability we further evaluate on the two multimodal sarcasm dataset MMSD[2] and MMSD2.0[18].

Table 2: Performance comparison on the MMSD and MMSD2.0 dataset.

Datasets	Methods	Acc	P	R	F1
MMSD	HFM[2]	0.834	0.766	0.842	0.802
	D&R Net[29]	0.840	0.780	0.834	0.806
	Att-BERT[16]	0.861	0.809	0.851	0.829
	InCrossMGs [8]	0.861	0.814	0.844	0.828
	HCM [11]	0.874	0.818	0.865	0.841
	DynRT-Net [26]	**0.936**	**0.931**	**0.936**	**0.933**
	LyRE(Ours)	0.901	0.899	0.897	0.898
MMSD2.0	HFM[2]	0.706	0.648	0.691	0.669
	HKE [10]	0.765	0.735	0.711	0.723
	CMGCN [9]	0.798	0.758	0.780	0.769
	Att-BERT[16]	0.800	0.763	0.778	0.770
	DynRT-Net [26]	0.714	0.718	0.722	0.713
	LyRE(Ours)	**0.816**	**0.764**	**0.829**	**0.795**

Table 3: Ablation study of LyRE on the Weibo-21 dataset.

	Evaluation Metric			
	Δ Acc	Δ F1	Acc	F1
Text	-4.17%	-4.23%	0.906	0.906
Image	-23.90%	-23.56%	0.708	0.713
Sum	-2.41%	-2.35%	0.923	0.925
Co-Attn	-2.19%	-2.18%	0.925	0.927
Concat	-1.64%	-1.62%	0.931	0.932
F	-1.10%	-1.17%	0.936	0.937
S	-1.75%	-1.80%	0.930	0.930
T	-0.77%	-0.80%	0.940	0.940
F+S	-0.99%	-1.00%	0.938	0.938
F+T	-0.32%	-0.36%	0.944	0.945
S+T	-0.65%	-0.66%	0.941	0.942
R	-1.21%	-1.26%	0.935	0.936
LyRE+R	-0.77%	-0.84%	0.940	0.940
w/o CF-CL	-0.86%	-0.72%	0.938	0.940
LyRE	-	-	0.947	0.948

As the first layer has 4 pathways, the second layer has 6, and the third layer has 9, we set $\alpha = \frac{9}{19}$, $\beta = \frac{6}{19}$, and $\gamma = \frac{4}{19}$. This weighting scheme ensures that the expected gradients from the three losses are relatively balanced.

3.2 Baseline Methods

To ensure a fair comparison, we selected multi-modality baseline methods with comparable backbone model performance. For the misinformation detection task, we compare the performance of LyRE against recent multi-modality methods: EANN [27], SpotFake [25], CAFE [4], SAFE [34], and BMR [30]. On the sarcasm detection dataset, we benchmark against: HFM [2], Att-BERT[16], CMGCN[9], HKE [10] and DynRT-Net [26].

3.3 Performance Comparison

Tables 1 and 2 present the comparative results for misinformation detection and sarcasm detection, respectively. We conduct a comprehensive performance evaluation using accuracy, precision, recall, and F1-score as metrics, reporting the average evaluation metrics across five different initial seeds.

Comparison on Misinformation Detection Employing a multi-output classification approach demonstrably enhances performance. While the baseline EANN model achieves accuracies of 86.4% and 87.0% on the GossipCop and Weibo-21 datasets, respectively, SpotFake exhibits comparatively lower performance, reaching 85.8% and 85.1% on the same datasets. The CAFE method, by refining the learning focus of the two output classifiers, further improves performance, achieving accuracies of 86.7%, 86.4%, and 88.2% on GossipCop, PolitiFact, and Weibo-21, respectively. Notably, the BMR model, which integrates multiple unimodal and multimodal perspectives, yields the highest accuracies,

reaching 89.7%, 88.5%, and 92.3% on GossipCop, PolitiFact, and Weibo-21, respectively.

LyRE demonstrates highly competitive performance, achieving accuracies of 89.7%, 94.2%, and 94.7% on the GossipCop, PolitiFact, and Weibo-21 datasets, respectively. These results compare favorably to state-of-the-art baseline BMR . Notably, our approach further improves upon the BMR baseline by 0.7%, 5.7%, and 2.4% on the respective datasets. Furthermore, LyRE consistently yields the highest F1-scores across all datasets, indicating a more balanced classification performance across different categories.

Comparison on Sarcasm Detection The DynRT-Net methodexhibited a significant performance discrepancy between the MMSD and MMSD2.0 datasets, achieving 96.6% accuracy on the former but only 71.4% on the latter. This discrepancy arises from data leakage issues within the MMSD dataset, where models could achieve artificially high performance by overfitting to spurious correlations in the text. MMSD2.0 addresses this leakage through rigorous text cleaning, forcing models to rely on genuine cross-modal interactions rather than relying solely on textual cues. Notably, our method achieves an accuracy of 81.4% on MMSD2.0, outperforming other methods with similar backbone model architectures. This result highlights the effectiveness of our approach in fusing visual and textual information to learn robust and generalizable cross-modal representations.

3.4 Abalation Study

To validate the contribution of each module in our proposed method, we conduct a series of ablation studies. "Text" and "Image" denote the use of unimodal features only. "Co-Attn", "Concat", and "Sum" represent replacing our proposed hierarchical aggregation module with co-attention, concatenation, and summation, respectively. "F", "S", and "T" correspond to the classifiers built upon the first, second, and third aggregation layers, respectively. We analyze variants with different layer combinations. To investigate the impact of connection depth, we experimented with adding a fourth aggregation layer, denoted by "R".

We investigated the impact of utilizing different combinations of fusion layers. As shown in 3,when using a single classifier head, the third layer "T" achieves the best performance, while the second layer "S" exhibits comparatively weaker results. When combining two classifier heads, the "F+T" configuration slightly outperformed "S+T" but demonstrated significantly better performance compared to "F+S." This suggests that the information learned by the first and second layers is relatively similar, resulting in a limited gain from their combination. In contrast, combining the first and third layers provides a more substantial information gain.

To explore the effects of fusion layer depth, we experimented with incorporating a fourth layer. Results indicate that the standalone fourth layer classifier "R" yielded inferior performance compared to the third layer "T". Similarly, combining all four layers "LyRE+R" led to a performance degradation compared

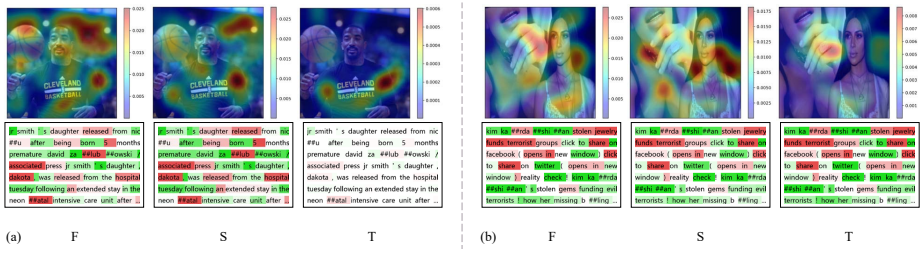


Fig. 4: Visualization of interpretability for strongly and weakly complementary multimodal samples. Deeper green highlights indicate stronger contributions to a "real" prediction, while deeper red highlights indicate stronger contributions to a "fake" prediction. (a) illustrates a weakly complementary case, while (b) depicts a strongly complementary one.

to using only the first three "F+T". This suggests that the increased complexity introduced by the fourth layer, where each unit receives connections from four pathways, might lead to overfitting and an excessive fusion of information, ultimately resulting in information loss.

4 Case Study

To analyze the distinct features learned by different hierarchical aggregation layers, we visualize the gradient influence of text and image modalities, as shown in Figure 4. When a significant semantic gap exists between modalities, as in Figure 4(a) where the text describes "*daughter of Jr Smith*" while the image portrays "*Jr Smith playing basketball*," the gradient magnitude at the highly-fused third layer is notably smaller than at the first layer. Conversely, in Figure 4(b), where both text and image relate to "*jewelry*," the third layer exhibits more precise attention focused on the ring on the hand. This suggests that each layer learns different levels of information fusion with varying attention foci, effectively performing detection in both low and high cross-modal complementarity scenarios.

5 Conclusion

In this paper, we introduce LyRE, a simple but effective framework for multimodal misinformation detection. LyRE comprises two main modules: Hierarchical Aggregation Network and Varying Fusion Degrees classifiers. The hierarchical aggregation mechanism uses a carefully designed pathway for progressive cross-modal fusion, allowing the model to learn from different levels of inter-modal interaction. This approach enables comprehensive learning from multimodal information and reduces information loss during cross-modal fusion, resulting in superior misinformation detection performance. Our experimental results explore the distinct characteristics learned at different fusion degrees, providing

valuable insights into how multimodal models effectively capture and leverage cross-modal correlations.

Acknowledgments. This work is supported by Tianshan Talent Training Program(Grant No. 2023TSYCCX0041), the Natural Science Foundation of Xinjiang Uyghur Autonomous Region(Grant No. 2022D01D81, Grant No.2022D01D 04), the Outstanding Member Program of the Youth Innovation Promotion Association of Chinese Academy of Sciences(Grant No. Y2023118, Y2021112), the Key Research and Development Program of Xinjiang Uyghur Autonomous Region(Grant No. 2024B03026, 2023B03024), the "Tianshan Elite" Science and Technology Topnotch Youth Talents Program(Grant No. 2022TSYCCX0059, 2023TSYCCX0044), the "Tianshan Elite" Science and Technology Innovation Leading Talents Program(Grant No. 2022TSYCLJ0046), the Young Scientists Fund of Natural Science Foundation of Xinjiang Uyghur Autonomous Region (Grant No. 2022D01B207) and the Youth Talents Support Project of Xinjiang Uyghur Autonomous Region(Grant No. 2023TSYCQNTJ0037).

References

1. Brady, W.J., Gantman, A.P., Van Bavel, J.J.: Attentional capture helps explain why moral and emotional content go viral. Journal of Experimental Psychology: General **149**(4), 746 (2020)
2. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2506–2515. Association for Computational Linguistics, Florence, Italy (Jul 2019)
3. Calvert, G.A., Thesen, T.: Multisensory integration: methodological approaches and emerging principles in the human brain. Journal of Physiology-Paris **98**(1-3), 191–205 (2004)
4. Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., Shang, L.: Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM web conference 2022. pp. 2897–2905 (2022)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
6. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
7. Li, P., Sun, X., Yu, H., Tian, Y., Yao, F., Xu, G.: Entity-oriented multi-modal alignment and fusion network for fake news detection. IEEE Transactions on Multimedia **24**, 3455–3468 (2021)
8. Liang, B., Lou, C., Li, X., Gui, L., Yang, M., Xu, R.: Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In: Proceedings of the 29th ACM international conference on multimedia. pp. 4707–4715 (2021)

9. Liang, B., Lou, C., Li, X., Yang, M., Gui, L., He, Y., Pei, W., Xu, R.: Multi-modal sarcasm detection via cross-modal graph convolutional network. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). vol. 1, pp. 1767–1777. Association for Computational Linguistics (2022)
10. Liu, H., Wang, W., Li, H.: Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 4995–5006. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022)
11. Liu, H., Wang, W., Li, H.: Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. arXiv:2210.03501 (2022)
12. McDonald, J.J., Teder-Sälejärvi, W.A., Ward, L.M.: Multisensory integration and crossmodal attention effects in the human brain. Science **292**(5523), 1791–1791 (2001)
13. Nan, Q., Cao, J., Zhu, Y., Wang, Y., Li, J.: Mdfend: Multi-domain fake news detection. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management. pp. 3343–3347 (2021)
14. Nielsen, D.S., McConville, R.: Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3141–3153. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022)
15. Osmundsen, M., Bor, A., Vahlstrup, P.B., Bechmann, A., Petersen, M.B.: Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. American Political Science Review **115**(3), 999–1015 (2021)
16. Pan, H., Lin, Z., Fu, P., Qi, Y., Wang, W.: Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1383–1392 (2020)
17. Pennycook, G., Rand, D.G.: The psychology of fake news. Trends in cognitive sciences **25**(5), 388–402 (2021)
18. Qin, L., Huang, S., Chen, Q., Cai, C., Zhang, Y., Liang, B., Che, W., Xu, R.: Mmsd2. 0: towards a reliable multi-modal sarcasm detection system. arXiv preprint arXiv:2307.07135 (2023)
19. Rao, Y., Zhao, W., Zhu, Z., Lu, J., Zhou, J.: Global filter networks for image classification. Advances in neural information processing systems **34**, 980–993 (2021)
20. Thiebaut de Schotten, M., Forkel, S.J.: The emergent properties of the connected brain. Science **378**(6619), 505–510 (2022)
21. Senkowski, D., Engel, A.K.: Multi-timescale neural dynamics for multisensory integration. Nature Reviews Neuroscience pp. 1–18 (2024)
22. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data **8**(3), 171–188 (2020)
23. Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake news detection on social media: A data mining perspective. ACM SIGKDD explorations newsletter **19**(1), 22–36 (2017)
24. Siddiqui, T.J.: Bridging the semantic gap (2015)
25. Singhal, S., Shah, R.R., Chakraborty, T., Kumaraguru, P., Satoh, S.: Spotfake: A multi-modal framework for fake news detection. In: 2019 IEEE fifth international conference on multimedia big data (BigMM). pp. 39–47. IEEE (2019)

26. Tian, Y., Xu, N., Zhang, R., Mao, W.: Dynamic routing transformer network for multimodal sarcasm detection. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2468–2480 (2023)
27. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., Gao, J.: 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. pp. 849–857 (2018)
28. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)
29. Xu, N., Zeng, Z., Mao, W.: Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In: Proceedings of the 58th annual meeting of the association for computational linguistics. pp. 3777–3786 (2020)
30. Ying, Q., Hu, X., Zhou, Y., Qian, Z., Zeng, D., Ge, S.: Bootstrapping multi-view representations for fake news detection. In: Proceedings of the AAAI conference on Artificial Intelligence. vol. 37, pp. 5384–5392 (2023)
31. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? Advances in neural information processing systems **27** (2014)
32. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)
33. Zheng, J., Zhang, X., Guo, S., Wang, Q., Zang, W., Zhang, Y.: Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In: IJCAI. pp. 2413–2419 (2022)
34. Zhou, X., Wu, J., Zafarani, R.: : Similarity-aware multi-modal fake news detection. In: Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II. p. 354–367. Springer-Verlag, Berlin, Heidelberg (2020)
35. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multimodal fake news detection via clip-guided learning. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 2825–2830. IEEE (2023)