HF-Mamba: Improving Multimodal Classification via Hierarchical Fusion based on Mamba

Yimo Ren, Jinfa Wang, Hong Li✉, Rongrong Xi, Haiqiang Fei, and Hongsong Zhu

Institute of Information Engineering,Chinese Academy of Sciences, Beijing, China
renyimo,wangjinfa,lihong,xirongrong,feihaiqiang,zhuhongsong@iie.ac.cn

Abstract. Multimodal fusion seeks to enhance the performance of models for various applications by extracting and integrating information from multiple modalities, such as text, images, and others. Recent studies have demonstrated the advantages of Transformer-based approaches for multimodal fusion in numerous multimedia tasks. However, Transformers often face efficiency challenges when handling long-range sequence modeling. In this paper, we address common multimodal classification tasks in social media, specifically sarcasm detection and sentiment analysis. We introduce HF-Mamba (Hierarchical Fusion based on Mamba), a novel framework designed to achieve superior multimodal fusion. HF-Mamba processes textual and visual data as sequences and leverages Mamba's capability for long-range sequence learning to integrate information across modalities at multiple levels effectively. Experiments conducted on two widely used public datasets from Twitter and Yelp validate the effectiveness of HF-Mamba. The results demonstrate that HF-Mamba can achieve state-of-the-art performance for sarcasm detection and sentiment analysis, outperforming existing baseline models.

Keywords: Multimodal fusion · State Space Model · Long Range Dependency.

1 Introduction

Multimodal fusion, the integration of information from multiple modalities such as text, image, and audio, holds significant importance across various real-world applications. In this paper, we focus on sarcasm detection and sentiment analysis, namely two multimodal classification tasks, which include texts and images. Current state-of-the-art multimodal fusion methods predominantly leverage Transformer-based architectures. The core of the Transformer-based architecture is the attention structure [1], which can learn the interactions between representations of different modalities. However, several challenges persist: (1)

Hong Li is the corresponding author.

Supported by Youth Innovation Promotion Association,Chinese Academy of Sciences(No.E3YY031104).

Existing methods necessitate the design of specific feature extractors tailored to individual modalities. This kind of approach can be time-consuming and may not generalize well across diverse datasets or applications. (2) Some approaches, such as MetaTransformer [2] and OneLLM [3], treat different modalities uniformly as sequences. However, as the length of sequences increases, particularly in scenarios like image processing where the sequences are derived from pixels, Transformer-based models easily encounter computational inefficiencies and performance degradation. Therefore, Transformer-based models could not learn the long range dependency of input sequences. (3) Multimodal data often exhibits hierarchical interactions at various levels. For existing methods with sequence data as input, simply matching input sequences may not capture the dependencies between different modalities.

Therefore, this paper proposes HF-Mamba, a Hierarchical Fusion based on Mamba with Long Range Dependency for Multimodal Classification Tasks. In the context of multimodal classification tasks, HF-Mamba can unify data from different modalities into long sequences. By leveraging Mamba's capability in long sequence learning and incorporating hierarchical interactions, HF-Mamba effectively learns the interactions between input sequences. HF-Mamba reduces the need for specialized feature extractors by transforming different modalities to sequences and enables efficient learning of interactions among different modalities by employing Mamba. Additionally, HF-Mamba introduces a hierarchical fusion mechanism to capture interactions at multiple levels to enhance the richness of multimodal fusion.

To sum up, our contributions are: (1)As far as we know, this paper is the pioneering use of Mamba to realize the fine-grained multimodal fusion by learn the interactions between long-range sequences without any pre-trained models. (2)HF-Mamba introduces a novel approach to realize multimodal fusion based on hierarchical and multiple Mamba blocks, facilitating effective fusion of text and image modalities. (3)This paper conducts a series of experiments to demonstrate HF-Mamba could achieve the best performance for sarcasm detection and sentiment analysis, which outperforms existing state-of-the-art baselines.

2 Methodology

As illustrated in Figure 1, our proposed method comprises a **Tokenizer** to transformer texts and images into sequences, several **Mamba Blocks** to learn representations from individual and crossed modalities. Therefore, the paper introduces HF-Mamba in three parts: Input Sequences, Mamba Blocks and Hierarchical Fusion.

2.1 Input Sequences

Inspired by MetaTransformer [2], the multimodal inputs of HF-Mamba, such as texts and images, are processed by the **Tokenizer** as sequences for further fusion and classification.

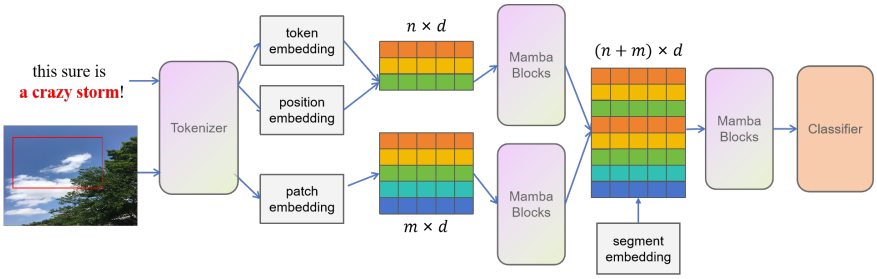

Fig. 1. The overview of proposed HF-Mamba.

For each input text, it is tokenized by WordPiece [4] into a sequence of tokens, denoted as $X = \{x_1, x_2, ..., x_n\}$, where n represents the number of tokens. Each token x_i is associated with a **token embedding** e_i. The dimension of e_i is set to d. To incorporate relative positional information, **position embeddings** pe_i are added to the token embeddings.

$$\begin{aligned} pe_{pos,2i} &= sin(\frac{pos}{10000^{\frac{2i}{d}}}) \\ pe_{pos,2i+1} &= cos(\frac{pos}{10000^{\frac{2i}{d}}}) \end{aligned} \tag{1}$$

Where pos represents the location of ith token, d is the dimension of the token embedding.

Therefore, the final token embedding t_i is obtained by summing the token embedding e_i and the corresponding position embedding pe_i:

$$t_i = e_i + pe_i \tag{2}$$

namely, the sequence embedding of text X is $T = \{t_1, t_2, ..., t_n\}$, $T \in R^{(n \times d)}$, where n represents the number of tokens from the text, d is the dimension of the sequence embedding.

For each input image, it is resized to the same size, for example 224×224. Then, a patching strategy is employed to divide the image into patches of a fixed size, such as 16×16. Each patch is treated as a text token, and patch embeddings are computed similar to text token embeddings. Therefore, each input image is tokenized into a sequence of tokens, denoted as $Y = \{y_1, y_2, ..., y_m\}$, where m represents the number of tokens. Each token y_i is associated with a **patch embedding** p_i. The dimension of p_i is the same as e_i. Therefore, the sequence embedding of image Y is $P = \{p_1, p_2, ..., p_m\}$, $P \in R^{(m \times d)}$, where m represents the number of tokens from the image, d is the dimension of the sequence embedding.

2.2 Mamba Block

State Space Models (SSMs) [5, 6] are a class of sequence-to-sequence modeling systems. These models, with linear complexity, could effectively capture the in-

herent dynamics of systems through an implicit mapping to latent states. A SSM could be defined as:

$$\begin{aligned} \dot{h}(t) &= Ah(t) + Bi(t) \\ o(t) &= Ch(t) + Di(t) \end{aligned} \tag{3}$$

where $i(t)$ and $o(t)$ are the inputs and outputs of SSM at time t, and $h(t)$ is the latent state of SSM. The parameters of SSM include $A \in C^{N \times N}$, $B, C \times C^N$ for a state size N, and the skip connection $D \in C^1$.

Subsequently, the process of discretization is commonly employed for practical deep learning algorithms. In this context, Δ represents the timescale parameter that is used to convert the continuous parameters A, B into discrete parameters, $\overline{A}, \overline{B}$. The zero-order hold method is commonly utilized for this discretization, and it is described as follows:

$$\begin{aligned} \overline{A} &= exp(\Delta A) \\ \overline{N} &= (\Delta A)^{-1}(exp(\Delta A) - I)\Delta B \end{aligned} \tag{4}$$

Once discretized, SSM can be simplified with the step size Δ as:

$$\begin{aligned} h_t &= \overline{A}h_{k-1} + \overline{B}i_k \\ o_t &= Ch_k + Di_k \end{aligned} \tag{5}$$

Nevertheless, due to the reason that parameters are invariant despite changes in the input, the formulation above is conducted on a Linear Time Invariance (LTI) system[7]. To address this constraint as much as possible, the recent work Mamba [8] explored integrating a selective scan technique to increase the model's capability by dynamically focusing on information from the input sequence.

In this situation, the matrices B, C, and Δ are derived from the input data. To simplify the relevant description, this paper sets that $i = \{i_1, i_2, ..., i_t\}$ represents the vectors of input sequences and $o = \{o_1, o_2, ..., o_t\}$ represents the vectors of output sequences.

Therefore:

$$o = SSM(i) \tag{6}$$

Then, a Simplified SSM architectures(Mamba) in [8] with i as inputs and o as outputs could be described as follows:

$$\begin{aligned} o = Linear(&SSM(Activation(Conv(Linear(i))))) \\ &+ Activation(Linear(i))) \end{aligned} \tag{7}$$

Where *Conv* is the One-dimensional convolution, *Linear* is the linear projection and *Activation* is the function of activation, such as SiLU function [9].

Therefore, for the embedding of input sequences X, the learned representation by Mamba blocks is as follows:

$$X \leftarrow Mamba(X) \tag{8}$$

2.3 Hierarchical Fusion

Hierarchical fusion can learn the matched information of multimodal data at different levels. To realize hierarchical fusion of texts and images, HF-Mamba firstly uses two separate Mamba blocks to learn texts and images sequences as follows:

$$\begin{aligned} T &\leftarrow Mamba_{text}(T) \\ P &\leftarrow Mamba_{image}(P) \end{aligned} \tag{9}$$

To incorporate the types of modalities, **segment embeddings** are added to the token embeddings of texts and images, which are represented as $Seg = \{seg_1, seg_2..., seg_k\}$.

For texts, the segment embedding is set to zeros with same dimension and same length of texts embedding. For images, the segment embedding set to zeros with same dimension and same length of images embedding. Therefore:

$$\begin{aligned} T &= T + Seg \\ P &= P + Seg \end{aligned} \tag{10}$$

Then, HF-Mamba concatenates the learnt texts and images sequences as a longer sequence for another Mamba block, as follows:

$$Z \leftarrow Mamba(T \oplus P) \tag{11}$$

where $\oplus$ represents the concatenation operation.

Therefore, the representation $Z = \{z_1, z_2, ..., z_{n+d}\} \in R^{(n+m)\times d}$ is obtained to train a classifier to realize sarcasm detection or sentiment analysis.

3 Experiments

3.1 Dataset

The paper utilizes two publicly available multimodal sentiment datasets, namely Twitter Sarcasm and Yelp Sentiment, both of which are built for classification tasks, to evaluate the proposed method HF-Mamba. Each dataset is divided into three parts: Training, Development and Testing. Every example in the dataset consists of a text and an associated image.

- **Twitter Sarcasm** is collected by [10], and it is a sentence-level dataset because the text is relatively short. This dataset contains English tweets expressing sarcasm labeled as 1 and those expressing non-sarcasm labeled as 0. The number of the training set, development set and testing set in Twitter sarcasm are 19816, 2410 and 2409, respectively. The average length of samples of Twitter sarcasm is about 15 words. Therefore, it is a sentence-level dataset for the classification tasks.

– **Yelp Sentiment** is built from Yelp restaurant reviews by [11], covering five different major US cities. It is a document-level dataset because the text is relatively long. The rating of the reviews in the dataset is used as the sentiment label in the range of 1-5. It should be noted that the number of each class in Yelp Sentiment is set to the same. The number of the training set and development set is 35435 and 2215. And the number of samples of the testing set in Boston (BO), Chicago (CH), Los Angeles (LA), New York (NY), and San Francisco (SF) is 315, 325, 3730, 1715, 570. The average length of samples of Yelp sentiment is about 220 words. Therefore, it is a document-level dataset for the classification tasks.

3.2 Parameters

In order to achieve better results of HF-Mamba and for a more fair comparison with baselines, the hyper-parameters of HF-Mamba are set after multiple experiments. Some parameters are as follows: (1) The maximum length of the input texts to 15 on Twitter dataset and 128 on Yelp dataset. (2) All images are resized into 224×224 and the fixed size of patching strategy is 16×16. (3) The number of Mamba blocks in three kinds of Mamba layers is 5. (4) The dimension of all embeddings is 64. (5) The batch size and epochs for training are 128 and 100 separately.

3.3 Comparisons

In this section, the paper compares performance between HF-Mamba and baseline models. For Twitter Sarcasm, the paper mainly reports the metrics $precision(P)$, $recall(R)$, and $F1$ of positive samples and $accuracy(acc)$ of total samples to show the ability to detect sarcasm. Because of the balance data of Yelp Sentiment, the paper reports $accuracy(acc)$ of total samples in different cities to show the performance of sentiment analysis.

Table 1 and Table 2 show the results of comparisons between HF-Mamba and baselines.

We find that almost all results of multimodal models are better than those of unimodal approaches, which validates again the importance of visual information for multimodal classification tasks. For example, the best of unimodal models is BERT, with 0.8410 accuracy and 0.5800 accuracy in Twitter Sarcasm and Yelp Sentiment, lower than the performance of HFM and VistaNet, respectively. Compared with unimodal methods, HF-Mamba achieves best performance on Twitter Sarcasm and Yelp Sentiment. HF-Mamba achieves 0.9191 accuracy and 0.6384 accuracy in Twitter Sarcasm and Yelp Sentiment. The results show that HF-Mamba can still realize fine multimodal fusion by converting data in different modalities into sequences. Compared with all methods, we could find the effectiveness of HF-Mamba on the sentence-level dataset Twitter Sarcasm and the document-level dataset Yelp Sentiment, respectively. HF-Mamba performs better at all metrics in Twitter Sarcasm and key metrics in Yelp Sentiment,

Table 1. The Comparisons between HF-Mamba and Baselines on Twitter Sarcasm.

Modality	Method	Twitter Sarcasm			
		Accuracy	Precision	Recall	F1
Text	TextCNN	0.8003	0.7429	0.7439	0.7532
	BiGRU	0.7978	0.7545	0.7288	0.7414
	HAN	0.7750	0.7299	0.6903	0.7095
	SIARN	0.8057	0.7555	0.7570	0.7563
	SMSD	0.8090	0.7646	0.7518	0.7582
	BERT	0.8410	0.7874	0.8227	0.8047
Text and Image	HFM	0.8344	0.7657	0.8415	0.8018
	D&R Net	0.8402	0.7797	0.8342	0.8060
	Res-BERT	0.8480	0.7780	0.8415	0.8085
	Att-BERT	0.8605	0.7863	0.8331	0.8090
	InCrossMGs	0.8610	0.8138	0.8436	0.8284
	CMGCN	0.8755	0.8363	0.8469	0.8416
	DMF-RHGT-HPA	0.9087	0.8927	0.8759	0.8842
	HF-Mamba	**0.9191**	**0.9210**	**0.8816**	**0.9009**

The results of baselines are retrieved from corresponding research[12].

Table 2. The Comparisons between HF-Mamba and Baselines on Yelp Sentiment.

Modality	Method	Yelp Sentiment					
		BO_Acc	CH_Acc	LA_Acc	NY_Acc	SF_Acc	Total_Acc
Text	TextCNN	0.5560	0.5540	0.5440	0.5420	0.5300	0.5430
	BiGRU	0.5494	0.5602	0.5645	0.5827	0.5280	0.5652
	HAN	0.6160	0.5850	0.5760	0.5710	0.5300	0.5730
	BERT	0.5714	0.6154	0.5778	0.5802	0.5789	0.5800
Text and Image	VistaNet	0.5840	0.6370	0.5900	0.5910	0.5510	0.5890
	GAFN	0.6160	**0.6620**	0.5900	0.6100	0.6070	0.6010
	VisdaNet	0.6286	0.6277	0.6257	0.6210	0.6070	0.6232
	DMF-RHGT-HPA	0.6031	0.6461	**0.6501**	0.6327	0.5930	0.6390
	HF-Mamba	**0.6317**	0.6338	0.6445	**0.6356**	**0.6140**	**0.6384**

The results of baselines are retrieved from corresponding research[12].

which are in bold in Table 1 and Table 2. Therefore, HF-Mamba is one of the state-of-the-art models on multimodal classification tasks to date.

Thus, HF-Mamba could achieve state-of-the-art performance on publicly available datasets for sarcasm detection and sentiment analysis. The evaluation of HF-Mamba proves that Mamba could be a possible choice for achieving better fusion of multimodal data.

3.4 Ablations

The core components in HF-Mamba to realize multimodal classification tasks are **Input Sequences**, **Mamba Blocks** and **Hierarchical Fusion**. The paper conducts ablation studies to show their effects further. All experiments of ablation are conducted on a GeForce RTX 3090 with PyTorch 1.13.0.

Input Sequences HF-Mamba converts data in different modalities into sequences and inputs them into Hierarchical Mamba blocks. Therefore, this paper explores the capability of Mamba blocks for multimodal fusion by varying their inputs, as Table 3 shows.

Table 3. The ablations for different inputs of HF-Mamba.

Text	Image	Accuracy	
		Twitter Sarcasm	Yelp Sentiment
✓		0.8024	0.5769
	✓	0.8481	0.6024
✓	✓	**0.9191**	**0.6384**

The results in the Table 3 tell that the information of images in the type of sequences are still helpful and could be learned well by HF-Mamba, so that HF-Mamba performs best when it both uses texts and images.

Mamba Blocks HF-Mamba leverages Mamba's advantages in learning long range dependency from the input sequences, enabling itself to attain strong ability to represent multimodal data even with simpler architectures.

Table 4. The ablations for different backbone of HF-Mamba.

Backbone	Accuracy	
	Twitter Sarcasm	Yelp Sentiment
Self-Attention(number of heads = 1)	0.5737	0.4902
Self-Attention(number of heads = 2)	0.5762	0.4476
Self-Attention(number of heads = 4)	0.5682	0.4523
Vanilla Transformer(number of heads = 1)	0.7335	0.4575
Vanilla Transformer(number of heads = 2)	0.8057	0.5025
Vanilla Transformer(number of heads = 4)	0.7547	0.4675
Mamba	**0.9191**	**0.6384**

To experimentally verify the assumption above, this paper replaces Mamba blocks with Self-Attention and Vanilla Transformer, which are designed in [1]. The results of experiments are shown in the Table 4. The results prove that Mamba surpasses Vanilla Transformer and Self-Attention in both time efficiency and performance, when implementing multimodal fusion through learning from long sequences of texts and images.

It should be noted that the current combination of HF-Mamba's blocks is relatively simple, so its performance is not as good as our intuitive expectation in the Table 4 if we simply replace Mamba blocks with Self-Attention or Vanilla Transformer.

Hierarchical Fusion Hierarchical fusion can learn the matched information of multimodal data at different levels. To prove this, this paper simply concatenates vectors of texts and images and inputs them into Mamba, which is called **Flatten-Mamba**. Also, this paper uses Mamba to extract features from input vectors of texts and images and concatenates them without using segment embedding, which is called **HF-Mamba-Simple**. The results of experiments are shown in the Table 5.

Table 5. The ablations for hierarchical fusion of HF-Mamba.

Type	Accuracy	
	Twitter Sarcasm	Yelp Sentiment
Flatten-Mamba	0.7609	0.4524
HF-Mamba-Simple	0.8455	0.5454
HF-Mamba	**0.9191**	**0.6384**

The results in the Table 5 indicate that through Hierarchical Fusion, Mamba is able to learn matching information of modalities at different levels, thereby achieving better multimodal fusion than that of Flatten-Mamba.

Further, the reason why HF-Mamba outperforms HF-Mamba-Simple is that segment embedding plays a directional role in different modalities in hierarchical fusion, which enhances the learning ability of HF-Mamba.

4 Conclusion

To improve the performance of classification tasks, such as sarcasm detection and sentiment analysis in social media, the paper proposes HF-Mamba, including Sequence Input, Mamba Block and Hierarchical Fusion. The paper conducts experiments on two primary and public datasets from Twitter and Yelp, respectively. The results show that our model could achieve the best performance for sarcasm detection and sentiment analysis, which outperforms existing state-of-the-art baselines. HF-Mamba leverages Mamba's advantages in learning long range dependency from the input sequences, enabling itself to attain strong ability to represent multimodal data even with simpler architectures. The evaluation of HF-Mamba proves that Mamba could be a possible choice for achieving better fusion of multimodal data.

Due to the ability of Mamba to contain many different types of modalities in the type of sequences, we are applying HF-Mamba to more classification tasks on multimodal graphs with more modalities. We believe our method could still obtain competitive performance in multimodal scenarios, including texts, images, and audios, *etc.*

References

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
2. Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023.
3. Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. *arXiv preprint arXiv:2312.03700*, 2023.
4. Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. Fast WordPiece tokenization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
5. Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.
6. Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.
7. Ramin Hasani, Mathias Lechner, Tsun-Hsuan Wang, Makram Chahine, Alexander Amini, and Daniela Rus. Liquid structural state-space models. *arXiv preprint arXiv:2209.12951*, 2022.
8. Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024.
9. Ningning Ma, Xiangyu Zhang, Ming Liu, and Jian Sun. Activate or not: Learning customized activation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8028–8038, 2021.
10. Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515, 2019.
11. Quoc-Tuan Truong and Hady W Lauw. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 305–312, 2019.
12. Yimo Ren, Jinfa Wang, Jie Liu, Peipei Liu, Hong Li, Hongsong Zhu, and Limin Sun. A relation-aware heterogeneous graph transformer on dynamic fusion for multimodal classification tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7855–7859. IEEE, 2024.