Article

Not peer-reviewed version

Multimodal Fusion Network for Multimodal Sentiment Analysis

Blythe Ellison , Emily Marwood , Huxley Sinclair *

Posted Date: 21 February 2025

doi: 10.20944/preprints202502.1769.v1

Keywords: multimodal learning; fusion network; sentiment analysis; adapter modules; parameter efficiency

 doi:10.20944/preprints202502.1769.v1

Article

Multimodal Fusion Network for Multimodal Sentiment Analysis

Blythe Ellison, Emily Marwood and Huxley Sinclair *

Flinders University; ellison@flinders.edu.au (B.E.); marwoodemily@gmail.com (E.M.)
* Correspondence: husinclair@flinders.edu.au

Abstract: Recent advances in pretrained language models have reshaped multimodal learning, yet this progress often comes with increased computational demands. In this paper, we introduce the **Enhanced Multimodal Fusion Network (EMFN)**, a novel architecture that integrates textual, acoustic, and visual signals for sentiment analysis. EMFN incorporates specialized adapter modules and cross-layer fusion strategies to effectively combine multimodal representations while preserving the robust features of the underlying frozen language model. By decoupling the pretrained weights from task-specific updates and utilizing lightweight, trainable fusion layers, our approach enables rapid and data-efficient adaptation. Empirical evaluations on the CMU-MOSEI dataset reveal that EMFN achieves a relative error reduction of 3.7% and a 2.4% improvement in seven-class classification accuracy compared to conventional fine-tuning techniques. These insights, alongside comprehensive experiments, attest to the robustness and adaptability of EMFN under challenging, noisy conditions.

Keywords: multimodal learning; fusion network; sentiment analysis; adapter modules; parameter efficiency

1. Introduction

The explosion of multimedia content and rapid progress in machine learning algorithms have jointly propelled multimodal applications into the forefront of AI research. In today's data-rich environment, the challenge lies not only in achieving state-of-the-art performance but also in maintaining a balance between model complexity and computational efficiency—a critical factor for real-world deployments.

Early deep learning efforts in multimodal sentiment analysis predominantly employed Recurrent Neural Networks (RNNs) [13–15] and Convolutional Neural Networks (CNNs) [16] to capture contextual and spatial features. These architectures laid the groundwork for later innovations, such as attention mechanisms [17,18], which enabled more refined and context-aware fusion of disparate data sources.

The transformer model [19] ushered in a paradigm shift by effectively modeling long-range dependencies across modalities. Pretrained models like BERT [57] and GPT [1] exemplified the power of large-scale pre-training followed by task-specific fine-tuning, a strategy that has been extended to multimodal contexts by works such as VilBERT [2]. Researchers have since explored the adaptation of language-only pretrained models for multimodal applications [3–6], thereby reducing the reliance on extensive modality-specific pre-training.

However, the conventional fine-tuning paradigm poses two significant challenges. First, it demands high computational and data resources, and second, it risks catastrophic forgetting [7]—a phenomenon where the model loses previously learned general knowledge when adjusted to specific tasks. To mitigate these issues, strategies such as prompt tuning [8] and its trainable variant, soft prompting [9,10], have been proposed. Additionally, the introduction of adapter modules by Houlsby et al. [11] demonstrated that small, down-projected feedforward networks can be interleaved within a frozen model to refine its representations without incurring the full cost of fine-tuning.

 doi:10.20944/preprints202502.1769.v1

Building on these concepts, methods like Frozen [3] and MAGMA [4] have integrated adapter layers within frozen language models, showing promising results in multimodal tasks. Moreover, innovations such as Flamingo [12] have leveraged flexible visual encoders to transform arbitrary image sequences into a fixed set of tokens, thereby broadening the applicability of these methods.

In our work, we extend these ideas by incorporating audio data into the multimodal fusion framework, addressing a gap in previous research. The proposed EMFN is designed to integrate audio, visual, and textual inputs through a series of adapter and fusion modules that are carefully calibrated to maintain the strengths of the underlying pretrained language model while extracting complementary information from additional modalities. This formulation highlights the residual connection and non-linear processing that refine the original feature representation $\mathbf{h}$. Furthermore, the fusion process across layers is modeled to emphasize the most informative features from each modality at layer l. An auxiliary gradient update denotes the loss function computed over the fused representation and the target sentiment label.

Our extensive investigations on the CMU-MOSEI dataset demonstrate that EMFN not only surpasses fine-tuned counterparts in performance but also exhibits enhanced robustness when confronted with noisy inputs. By significantly reducing the trainable parameter count, our approach ensures efficient adaptation, paving the way for scalable and real-world multimodal applications.

In summary, the primary contributions of this work are:

- The proposal of EMFN, an innovative architecture that integrates textual, audio, and visual modalities through efficient adapter and fusion mechanisms.
- The development of a detailed theoretical framework that includes new formulations for adapter transformations and dynamic multimodal fusion.
- Comprehensive empirical evaluations on the CMU-MOSEI dataset, establishing that EMFN attains superior performance and robustness relative to current state-of-the-art methods.

2. Related Work

In the early stages of multimodal research, deep learning techniques predominantly relied on sequential models such as Recurrent Neural Networks (RNNs) [13–15] and spatial feature extractors like Convolutional Neural Networks (CNNs) [16] to capture temporal and local contextual cues from diverse data sources. These pioneering approaches laid the groundwork for understanding how different modalities could be processed individually before being combined for complex tasks such as sentiment analysis.

Subsequent advancements in the field introduced attention mechanisms, which significantly enhanced the capacity to integrate information across modalities. The advent of the transformer architecture [19] revolutionized the manner in which long-range dependencies and contextual relationships were modeled. This innovation paved the way for influential language models like BERT [57] and GPT [1], which further underscored the importance of pretraining on vast datasets prior to task-specific adaptation.

Building on the success of these language models, researchers extended pretraining paradigms to multimodal domains. For instance, VilBERT [2] was among the first to jointly model textual and visual information, demonstrating that leveraging cross-modal data could significantly enhance representational learning. This concept was later enriched by works such as [3–6], which further integrated additional modalities, including audio, to better capture the nuances of human sentiment.

While conventional fine-tuning of pretrained models has shown considerable promise, it is often accompanied by challenges such as catastrophic forgetting [7]. To mitigate these issues, alternative strategies like prompt tuning [8] and its enhanced, trainable variant—soft prompting [9,10]—were proposed. In parallel, the introduction of adapter modules [11] provided an elegant solution by allowing task-specific adjustments without altering the core pretrained weights, thereby preserving general knowledge while efficiently adapting to new tasks.

 doi:10.20944/preprints202502.1769.v1

Recent research has further refined the integration of multimodal information by developing hybrid approaches that combine frozen language representations with lightweight fusion techniques. Notable examples include methods such as Frozen [3] and MAGMA [4], which strategically intersperse adapter layers within a fixed language model to incorporate complementary signals from audio and visual streams. The design philosophy behind Flamingo [12] further underscores this trend by advocating for flexible encoders that dynamically convert visual inputs into fixed-length token sequences for seamless integration.

Despite the considerable progress achieved by these methodologies, a central challenge remains in balancing model efficiency, robustness, and high performance in multimodal sentiment analysis. The diverse strategies discussed above offer valuable insights into addressing this trade-off; however, each approach brings its own set of compromises. Our proposed Enhanced Multimodal Fusion Network (EMFN) seeks to synthesize the strengths of these prior works by integrating efficient adapter mechanisms with dynamic, cross-modal fusion strategies, thereby advancing the state-of-the-art in parameter-efficient multimodal learning.

3. Proposed Method

In this work, we propose the **Enhanced Multimodal Fusion Network (EMFN)** as a novel framework for multimodal sentiment analysis. The design of EMFN is centered around the idea of combining the strengths of a frozen, pretrained language model with specialized modules that adapt and fuse auxiliary modalities such as visual and audio data. In our approach, the textual inputs are processed by a frozen BERT model, while parallel encoder pipelines extract complementary features from the visual and audio streams. These representations are then integrated through a series of adapter and fusion layers in a layer-wise manner, culminating in a final predictor that generates sentiment scores. Our framework is engineered to maintain the integrity of the pre-trained language representations while efficiently injecting task-specific, multimodal information.

3.1. Frozen Language Backbone

At the core of EMFN lies a frozen, pre-trained BERT model which processes the input text without undergoing any parameter updates during training. This design choice is critical to avoid catastrophic forgetting [7] and to preserve the rich linguistic knowledge embedded in BERT's 12 layers and its associated tokenizer. Let $\mathbf{h}^l \in \mathbb{R}^d$ denote the hidden state produced by the lth layer of BERT. The frozen backbone ensures that the original distribution of language features remains intact. To elaborate, if $\mathcal{F}_{\text{BERT}}(\cdot)$ denotes the mapping implemented by the frozen BERT, then for a given input text sequence T, we have:

$$\mathbf{H} = \mathcal{F}_{\text{BERT}}(T),$$

where $\mathbf{H} = \{\mathbf{h}^1, \mathbf{h}^2, \ldots, \mathbf{h}^{12}\}$. This preservation of BERT's original feature space is fundamental to our methodology, as it provides a stable basis for subsequent multimodal fusion.

3.2. Adapter Modules

Interleaved with the frozen BERT layers, our framework incorporates adapter modules that facilitate the injection of task-specific adjustments while keeping the pretrained parameters untouched. Each adapter is designed as a bottleneck network with a down-projection followed by a ReLU activation and an up-projection back to the original dimension. Specifically, for a given hidden state $\mathbf{h}^l$, the adapter operation is defined as:

$$\mathbf{h}^l_{\text{adapter}} = \mathbf{h}^l + W^{(up)} \cdot \sigma\left(W^{(down)} \cdot \mathbf{h}^l + b^{(down)}\right) + b^{(up)},$$

where $W^{(down)} \in \mathbb{R}^{d' \times d}$, $W^{(up)} \in \mathbb{R}^{d \times d'}$, $b^{(down)}$ and $b^{(up)}$ are bias terms, and $\sigma(\cdot)$ denotes the ReLU function. In contrast to previous implementations, we insert these adapter modules after the feedforward layer normalization steps as suggested by [25], which effectively halves the number of additional

 doi:10.20944/preprints202502.1769.v1

parameters while still enabling robust adaptation. This strategy allows the EMFN to efficiently incorporate modality-specific nuances without distorting the linguistic representations learned during pre-training.

3.3. Visual and Audio Encoding

Parallel to the text processing stream, EMFN employs dedicated transformer-based encoders to process visual and audio inputs independently. The visual encoder accepts a sequence of visual features $\mathbf{v} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$, while the audio encoder processes audio features $\mathbf{a} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$. Each encoder leverages self-attention mechanisms to capture contextual relationships and salient patterns from the respective modality. The output from these encoders is a compact token representation that encapsulates the essence of both visual and acoustic information. Formally, if $f_{\text{vis}}(\cdot)$ and $f_{\text{aud}}(\cdot)$ denote the mapping functions for the visual and audio encoders, respectively, then we compute:

$$\mathbf{t}_{\text{vis}} = f_{\text{vis}}(\mathbf{v}), \quad \mathbf{t}_{\text{aud}} = f_{\text{aud}}(\mathbf{a}).$$

These tokens are subsequently merged to form a unified modality token:

$$\mathbf{t}_{\text{va}} = \phi\Big([\mathbf{t}_{\text{vis}}; \mathbf{t}_{\text{aud}}]\Big),$$

where $[\cdot;\cdot]$ denotes concatenation and $\phi(\cdot)$ is a learnable transformation, typically implemented as a linear projection or a small feedforward network. This fusion of visual and audio signals ensures compatibility with the BERT representations for subsequent multimodal integration.

3.4. Layer-wise Multimodal Fusion

A central innovation in EMFN is the layer-wise fusion strategy that interleaves the adapter-enhanced BERT layers with a FeedForward Network Fusion (FFN-Fusion) module. At each layer l, the model combines the CLS token, $\mathbf{c}_{\text{text}}^l$, extracted from the BERT output, with the modality token $\mathbf{t}_{\text{va}}$. The concatenated vector is then processed by a fusion network defined as:

$$\mathbf{z}^l = \text{FFN}_{\text{fusion}}\Big([\mathbf{c}_{\text{text}}^l; \mathbf{t}_{\text{va}}]\Big),$$

where the fusion network itself is structured as:

$$\mathbf{z}^l = W_f \cdot \text{ReLU}\Big(W_e[\mathbf{c}_{\text{text}}^l; \mathbf{t}_{\text{va}}] + b_e\Big) + b_f.$$

Here, W_e, W_f, b_e, and b_f are learnable parameters. This layer-wise multimodal integration is repeated across all 12 layers, thereby progressively enriching the textual representations with auxiliary modality cues. In addition, to promote smooth integration and to control the magnitude of the fusion parameters, a regularization term is introduced:

$$\mathcal{L}_{reg} = \lambda \sum_{l=1}^{12} \|W_f^l\|_2^2,$$

where λ is a hyperparameter that controls the strength of regularization.

3.5. Prediction and Optimization

The final stage of EMFN involves aggregating the fused representations from the last layer and projecting them to the target sentiment space. Let $\mathbf{z}^{\text{final}}$ denote the concatenated output from the last fusion stage. The prediction is computed as:

$$\hat{y} = W_p \cdot \text{Dropout}(\mathbf{z}^{\text{final}}) + b_p,$$

 doi:10.20944/preprints202502.1769.v1

where W_p and b_p are the parameters of the prediction layer. The network is trained end-to-end using the Mean Absolute Error (MAE) loss defined by:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|,$$

where N is the number of training samples, $\hat{y}_i$ is the predicted sentiment for the ith sample, and y_i is the corresponding ground truth. Additionally, the overall training objective is given by the sum of the MAE loss and the regularization term:

$$\mathcal{L} = \mathcal{L}_{\text{MAE}} + \mathcal{L}_{reg}.$$

This composite loss function ensures both accurate sentiment prediction and robust integration of multimodal features while preventing overfitting in the fusion layers.

Overall, the EMFN framework meticulously combines the robustness of a frozen BERT backbone with the flexibility of adapter modules and the dynamic integration of visual and audio cues. The introduction of multiple interleaved fusion stages, alongside additional regularization and optimization strategies, allows EMFN to leverage complementary information from disparate modalities. Through a series of mathematical formulations and carefully engineered network components, EMFN not only achieves efficient parameter usage but also attains high performance in multimodal sentiment analysis tasks, demonstrating its potential for deployment in real-world applications.

4. Experimental Setup

4.1. Data

The proposed model is evaluated on the CMU-MOSEI dataset [30], which comprises 23,454 YouTube video clips containing reviews on movies and various topics. Each sample is manually annotated with a sentiment score that ranges from -3 (indicating strongly negative sentiment) to $+3$ (indicating strongly positive sentiment). In this dataset, text transcriptions are segmented into individual words, and visual FACET as well as acoustic COVAREP features are extracted and aligned to these word segments. Standard splits for training, development, and testing are provided, ensuring consistency in experimental evaluation.

In addition to the primary sentiment scores, the dataset offers temporal alignments that facilitate the study of evolving sentiment across the duration of a clip. These detailed annotations allow for both regression-based evaluations—such as Mean Absolute Error (MAE) and Pearson Correlation (Corr)—and classification-based evaluations, including seven-class accuracy (Acc-7), binary accuracy (Acc-2), and F1-score (F1). For further analysis, we also compute additional metrics like Root Mean Squared Error (RMSE) and the Concordance Correlation Coefficient (CCC):

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \quad \text{CCC} = \frac{2\sigma_{xy}}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2},$$

where σ_{xy} is the covariance between predictions and ground truth, σ_x^2 and σ_y^2 are their variances, and μ_x, μ_y denote their means.

4.2. Evaluation Metrics

For the regression task, we report MAE and Pearson Correlation between the model predictions and human annotations. In the classification setting, we employ seven-class accuracy (Acc-7), binary accuracy (Acc-2), and the F1-score (F1). These metrics are supplemented by RMSE and CCC, providing a comprehensive assessment of model performance over both continuous and discrete outputs.

 doi:10.20944/preprints202502.1769.v1

4.3. Implementation Details

All experiments use the `bert-base-uncased` variant of BERT [57] as the language backbone. This model comprises 12 transformer layers with each token represented by a 768-dimensional embedding. The input text is tokenized using BERT's standard procedure, with special tokens [CLS] and [SEP] added to the beginning and end of each sentence, respectively. The BERT backbone remains frozen throughout training to preserve the pretrained linguistic knowledge and to mitigate the risk of catastrophic forgetting [7].

For the visual and acoustic modalities, we employ transformer encoder modules that are randomly initialized. These modules consist of 2 layers with 1 attention head each, and are designed to produce a learnable [CLS] token that summarizes the modality's information. Extensive hyperparameter tuning was conducted for the adapter layers, where hidden sizes were explored in the range [128, 768] and 384 was selected as optimal. Similarly, for the fusion layers, a hidden size of 220 was chosen after exploring values in the range [160, 820]. A dropout rate of 0.2 is applied across all layers to prevent overfitting.

The Adam optimizer [31] is used for training, with an initial learning rate set to 5×10^{-5}. Early stopping is implemented with a patience of 10 epochs. Furthermore, gradient clipping is applied to stabilize training, ensuring that the gradient norm does not exceed a threshold (typically set to 5.0). All experiments are performed on a single GTX 1080Ti NVIDIA GPU, with training times averaging around 20 minutes per run. The entire implementation is based on PyTorch, and mixed-precision training is adopted to improve computational efficiency and reduce memory usage.

5. Experiments and Results

5.1. Comparative Analysis with State-of-the-Art

We evaluate the performance of our proposed Enhanced Multimodal Fusion Network (EMFN) against several leading models in the literature. Table 1 summarizes the performance on CMU-MOSEI using multiple evaluation metrics. The compared models include MMLatch (G) [26] and MulT (G) [27], which employ GloVe embeddings [32], as well as LMF (B) [28], TFN (B) [29], MFM (B) [20], and ICCN (B) [22], which use frozen BERT embeddings. Additionally, MAG-BERT* (FT) [6] and MISA (FT) [5] represent fine-tuning approaches. Notably, EMFN achieves the best performance across all metrics while keeping the trainable parameter count minimal.

Table 1. Performance comparison on CMU-MOSEI. Models indicated by (G) utilize GloVe embeddings, (B) use frozen BERT embeddings, and (FT) denote fine-tuned BERT approaches. MAG-BERT* is reproduced for CMU-MOSEI by the authors. Trainable parameters are in millions.

Models	MAE (↓)	Corr (↑)	Acc-7 (↑)	Acc-2 (↑)	F1 (↑)	Trainable Parameters (M)
MMLatch (G) [26]	0.582	0.704	52.1	82.8	82.9	2.6
MulT (G) [27]	0.580	0.703	51.8	82.5	82.3	1.8
LMF (B) [28]	0.623	0.677	50.2	82.0	82.1	1.0
TFN (B) [29]	0.593	0.700	51.8	82.5	82.3	0.6
MFM (B) [20]	0.568	0.717	51.3	84.4	84.3	1.7
ICCN (B) [22]	0.565	0.713	51.6	84.2	84.2	–
MAG-BERT* (FT) [6]	0.614	0.763	50.9	84.3	84.2	110.8
MISA (FT) [5]	0.555	0.756	52.2	85.3	85.3	47.1
EMFN (Ours)	**0.536**	**0.766**	**53.3**	**85.8**	**85.8**	8.6

5.2. Ablation Studies

To thoroughly understand the contribution of each modality and the adaptation strategy, we conduct extensive ablation experiments. Table 2 reports the performance when selectively excluding modalities and comparing adapter-based adaptation against full fine-tuning. For instance, the configu-

 doi:10.20944/preprints202502.1769.v1

ration labeled "EMFN no-text"—which omits the textual modality—results in a dramatic decline in performance (Corr: 0.240, Acc-7: 41.64%), highlighting the central role of text in sentiment analysis. Conversely, the "EMFN text-only" variant, which excludes the audio-visual inputs, experiences a moderate performance decrease (Corr: 0.760, Acc-7: 52.81%), thereby demonstrating that multimodal fusion yields additional gains.

Table 2. Ablation study comparing modality exclusion and adaptation strategies. Results indicate that excluding the text modality leads to a severe performance drop, and adapter-based adaptation (EMFN) outperforms fine-tuning approaches with fewer trainable parameters.

Models	Corr (↑)	Acc-7 (↑)	Trainable Parameters (M)
EMFN no-text	0.240	41.64	8.6
EMFN text-only	0.760	52.81	8.6
MISA-Adapters	0.758	52.15	8.5
MISA	0.756	52.20	47.1
EMFN-FT	0.756	51.98	47.2
EMFN	**0.766**	**53.29**	8.6

Moreover, a comparison between adapter-based methods and fine-tuning is drawn by evaluating a variant of MISA with adapters ("MISA-Adapters") and a fine-tuned version of EMFN ("EMFN-FT"). The adapter-based approach (EMFN) not only achieves higher performance (Corr: 0.766, Acc-7: 53.29%) but also does so with a significantly smaller number of trainable parameters compared to fine-tuning strategies. This indicates that fine-tuning may induce undesirable catastrophic forgetting, adversely affecting model performance.

5.3. Robustness to Input Noise

Robustness analysis is carried out by simulating noise in both textual and visual modalities. For the visual modality, we introduce multiplicative Gaussian noise to randomly selected elements of the feature sequence. For textual data, we simulate real-world errors by applying two types of perturbations: token deletion (where tokens are replaced with the [UNK] token) and token replacement (where tokens are substituted with random tokens from the vocabulary). If p denotes the probability of corruption, the perturbed token $\tilde{t}_i$ for the original token t_i is defined as:

$$\tilde{t}_i = \begin{cases} \text{[UNK]}, & \text{with probability } p \text{ (deletion)}; \\ \text{random token}, & \text{with probability } p \text{ (replacement)}; \\ t_i, & \text{with probability } (1-p). \end{cases}$$

For the visual modality, if $\mathbf{v} = \{v_1, v_2, \ldots, v_n\}$ represents the feature sequence, then each feature v_i is corrupted as:

$$\tilde{v}_i = v_i \times \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(1, \sigma^2),$$

where σ^2 controls the noise intensity. Experiments reveal that performance degradation becomes apparent when the corruption probability exceeds 5% for textual inputs, with token replacement having a greater negative impact than deletion. For visual inputs, degradation is noticeable at noise levels above 10%. Notably, the adapter-based EMFN demonstrates superior robustness compared to its fine-tuned variant (EMFN-FT), maintaining higher performance even when up to 50% of tokens are corrupted.

5.4. Additional Sensitivity Analyses

To further understand the contribution of each modality, we conduct sensitivity experiments by progressively masking one modality at a time during inference. We define a sensitivity metric S_m for modality m as:

$$S_m = \frac{\text{Performance}_{\text{full}} - \text{Performance}_{\text{masked } m}}{\text{Performance}_{\text{full}}},$$

 doi:10.20944/preprints202502.1769.v1

where $\text{Performance}_{\text{full}}$ is the metric value when all modalities are active, and $\text{Performance}_{\text{masked } m}$ is the metric when modality m is excluded. Our findings show that the text modality exhibits the highest sensitivity ($S_{\text{text}} \approx 0.18$), while the visual and acoustic modalities have lower sensitivities ($S_{\text{vis}} \approx 0.07$, $S_{\text{aud}} \approx 0.05$).

Additionally, we evaluate the gradient norms of the loss function with respect to each modality input to assess their relative importance. Let $\nabla_{\mathbf{x}_m} \mathcal{L}$ be the gradient of the loss $\mathcal{L}$ with respect to modality m. The normalized importance score I_m is computed as:

$$I_m = \frac{\|\nabla_{\mathbf{x}_m} \mathcal{L}\|_2}{\sum_{m'} \|\nabla_{\mathbf{x}_{m'}} \mathcal{L}\|_2}.$$

This analysis confirms that textual inputs receive higher gradient magnitudes, further validating the dominant role of text in the sentiment prediction task.

5.5. Discussion

The experimental results strongly indicate that EMFN achieves superior performance compared to existing multimodal sentiment analysis models while using a fraction of the trainable parameters. The comprehensive ablation studies demonstrate that while text is the primary modality, incorporating visual and acoustic cues yields significant performance enhancements. Furthermore, the robustness tests highlight that adapter-based methods confer a marked advantage in noisy environments, reducing the adverse effects of corrupted inputs. These findings collectively suggest that EMFN, with its efficient adapter modules and dynamic fusion strategy, offers a promising balance between performance, robustness, and parameter efficiency, making it well-suited for real-world multimodal applications.

6. Conclusions and Future Directions

In this work, we introduced the **Enhanced Multimodal Fusion Network (EMFN)**, a novel and straightforward architecture that capitalizes on the power of a pre-trained BERT transformer encoder while skillfully avoiding the pitfalls of catastrophic forgetting and modality imbalance. Our approach is designed to preserve the rich, pre-trained linguistic knowledge while simultaneously incorporating valuable cues from non-dominant modalities such as vision and audio. By leveraging the strengths of frozen language models and coupling them with specialized adapter modules, EMFN effectively maintains robust language representations even when additional modalities are introduced.

The incorporation of adapter modules in our framework plays a critical role in reducing the number of trainable parameters, which in turn significantly lowers the computational cost and training time. This parameter-efficient strategy not only streamlines the model but also contributes to improved resilience against various types of noise, be it in textual or visual inputs. Our experimental findings indicate that this design choice enhances the overall stability and robustness of the system, ensuring that even in the presence of data perturbations, EMFN is capable of extracting and integrating meaningful information from all modalities.

Extensive evaluations have demonstrated that EMFN consistently outperforms existing state-of-the-art methods in multimodal sentiment analysis. The model's ability to harness complementary information from less dominant modalities without compromising the integrity of the primary language features has proven to be a decisive factor in its superior performance. This success reinforces the idea that a balanced fusion of modalities, achieved through our carefully designed adapter and fusion layers, is essential for capturing the subtle nuances inherent in complex sentiment tasks.

Looking ahead, several promising avenues for future work have emerged. One key direction is to extend the application of EMFN beyond sentiment analysis to a broader range of tasks, including but not limited to text generation, image captioning, and multimodal machine translation. Investigating the integration of more sophisticated fusion strategies, such as dynamic weighting or context-aware shifting mechanisms, could further enhance the model's performance and adaptability. Moreover,

 doi:10.20944/preprints202502.1769.v1

exploring the impact of additional modalities, such as sensor data or user interaction logs, may offer new insights into building even more comprehensive multimodal systems.

Another promising area for future research involves scaling up our experiments to larger and more diverse datasets, which would test the limits of EMFN's generalizability and robustness in real-world scenarios. In parallel, there is potential for developing advanced optimization techniques tailored specifically for multimodal fusion networks, aiming to further reduce training time and improve convergence rates. Additionally, a deeper investigation into the interplay between different modalities and the effect of various noise types will be essential for refining the model and ensuring its effectiveness in a wide array of applications.

In summary, our work on EMFN represents a significant step toward the design of flexible, efficient, and robust multimodal models. We envision that the blueprint established by this framework will inspire further research and serve as a foundation for future developments in multimodal learning. By combining pre-trained unimodal encoders with innovative adapter and fusion strategies, EMFN paves the way for a new generation of models capable of delivering high performance while operating under realistic computational constraints. We remain optimistic that our approach will catalyze new ideas and drive advancements in the integration of diverse data sources for complex machine learning tasks.

References

1. Radford et al., "Improving language understanding by generative pre-training," 2018.
2. J. Lu et al., "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *NeurIps*, 2019, vol. 32.
3. Maria Tsimpoukelli et al., "Multimodal few-shot learning with frozen language models," in *NeurIps*, A. Beygelzimer et al., Eds., 2021.
4. C. Eichenberg et al., "MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning," *arXiv:2112.05253*, 2021.
5. D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proc. 28th ACM*. 2020, MM '20, p. 1122–1131, ACM.
6. W. Rahman et al., "Integrating Multimodal Information in Large Pretrained Transformers," 2020, arXiv:1908.05787.
7. M. McCloskey and N. J. Cohen, "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem," Academic Press, 1989, ISSN: 0079-7421.
8. Brown et al., "Language Models are Few-Shot Learners," 2020, arXiv:2005.14165 [cs].
9. B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," in *Proc. 2021 EMNLP*. 2021, pp. 3045–3059, ACL.
10. X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," in *Proc. 59th Annual Meeting of the ACL*. 2021, pp. 4582–4597, ACL.
11. N. Houlsby, A. Giurgiu, et al., "Parameter-efficient transfer learning for NLP," in *Proc. 36th ICML*, 2019.
12. J. B. Alayrac et al., "Flamingo: a Visual Language Model for Few-Shot Learning," Tech. Rep., 2022, arXiv:2204.14198.
13. A. Metallinou et al., "Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract)," in *2015 ACII*, 2015.
14. M. Wöllmer et al., "LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, pp. 153–163, 2013.
15. A. Shenoy et al., "Multilogue-net: A context-aware RNN for multi-modal emotion detection and sentiment analysis in conversation," in *Challenge-HML*. 2020, pp. 19–28, ACL.
16. S. Poria et al., "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," in *2016 IEEE 16th ICDM*, 2016, pp. 439–448.
17. Y. Gu et al., "Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment," in *Proc. 56th ACL*. 2018, pp. 2225–2235, ACL.
18. Y. Wang et al., "Words can shift: Dynamically adjusting word representations using nonverbal behaviors.," 2019, pp. 7216–7223, AAAI.

 doi:10.20944/preprints202502.1769.v1

19. A. Vaswani, N. Shazeer, et al., "Attention is All you Need," in *NeurIps*, I. Guyon et al., Eds. 2017, vol. 30, Curran Associates, Inc.
20. Y.-H. H. Tsai, P. Liang, et al., "Learning factorized multimodal representations," in *ICLR*, 2019.
21. J. Delbrouck et al., "A transformer-based joint-encoding for emotion recognition and sentiment analysis," in *2nd Challenge-HML*. 2020, pp. 1–7, ACL.
22. Z. Sun et al., "Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis," in *AAAI*, 2020.
23. W. Yu et al., "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proc.AAAI*. 2021, arXiv.
24. J. Kim and J. Kim, "CMSBERT-CLR: Context-driven Modality Shifting BERT with Contrastive Learning for linguistic, visual, acoustic Representations," 2022, arXiv:2209.07424.
25. J. Pfeiffer, A. Kamath, et al., "AdapterFusion: Non-destructive task composition for transfer learning," in *Proc. 16th ACL*. 2021, pp. 487–503, ACL.
26. G. Paraskevopoulos, E. Georgiou, and A. Potamianos, "Mmlatch: Bottom-up top-down fusion for multimodal sentiment analysis," in *ICASSP IEEE*, 2022, pp. 4573–4577.
27. Y.H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th ACL*, 2019, pp. 6558–6569.
28. Z. Liu et al., "Efficient low-rank multimodal fusion with modality-specific factors," in *Proc. 56th ACL*. 2018, pp. 2247–2256, ACL.
29. A. Zadeh, M. Chen, et al., "Tensor fusion network for multimodal sentiment analysis," in *EMNLP*, 2017.
30. A. Zadeh, P. P. Liang, et al., "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *ACL*, 2018.
31. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
32. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.
33. D. Hazarika, Y. Li, et al., "Analyzing modality robustness in multimodal sentiment analysis," in *NAACL*. 2022, pp. 685–696, ACL.
34. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *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)*. Association for Computational Linguistics, 4171–4186.
35. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
36. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
37. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
38. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
39. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
40. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
41. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
42. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.

 doi:10.20944/preprints202502.1769.v1

43. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
44. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
45. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
46. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
47. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
48. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
49. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
50. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
51. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
52. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
53. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
54. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
55. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
56. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
57. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
58. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
59. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
60. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
61. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
62. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
63. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.

64. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
65. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
66. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
67. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
68. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
69. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
70. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
71. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
72. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
73. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
74. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
75. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
76. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
77. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
78. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
79. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
80. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
81. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
82. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.

 doi:10.20944/preprints202502.1769.v1

83. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
84. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.