Journal of Digital Media & Culture Technology Vol.6, No.1, June (2026), pp.45-58
ISSN 2983-0664(p) 2983-0850(e) *http://dx.doi.org/10.29056/jdmct.2026.06.05*

Entity-Aware Bidirectional Attention-Based Gated Fusion for News Classification

Jun-Yeong Kang[1], Yousung Yeon[2], Chang Choi[3]*

Abstract

As the online news environment shifts toward combining diverse media such as text and images, research on multimodal AI-based news classification is underway to address the limitations of information loss in single-modality methods and improve classification accuracy by analyzing complex contexts. However, existing fusion techniques based on simple concatenation or unidirectional interaction struggle to preserve the core context in lengthy texts and fail to resolve the semantic discrepancies caused by irrelevant visual noise, ultimately leading to model bias and degraded classification performance. This paper proposes an entity-aware bidirectional attention-based gated fusion designed to maintain the key textual context and mitigate the bias issues induced by visual noise. The proposed architecture consists of a preprocessing stage that prevents context loss using NER and keywords extracted from unstructured text, a bidirectional cross-attention stage that aligns features through cross-modal referencing, and a gated fusion stage that suppresses noise amplification via modality-specific bias initialization. Experimental results using the New York Times N24News dataset demonstrate that the proposed model improves accuracy by up to 14.04% and F1-score by up to 15.73% compared to existing baseline models, validating its applicability for robust news classification even in environments characterized by semantic misalignment and noise.

Keyword : Multimodal News Classification, Bidirectional Cross-Attention, Gated Fusion, Named Entity Recognition, Visual Noise

1. Introduction

Multimodal learning combines and processes different modalities, such as text, images, audio, and video. Multimodal artificial intelligence (AI) is an AI technology that learns interrelationships between

1 Department of Computer Engineering, Gachon University, Seongnam, Korea [Undergraduate Student]
e-mail: jy0625@gachon.ac.kr

2 Department of Computer Engineering, Gachon University, Seongnam, Korea [Graduate Student]
e-mail: yus0613@gachon.ac.kr

3 Department of Computer Engineering, Gachon University, Seongnam, Korea [Professor]
e-mail: changchoi@gachon.ac.kr (Corresponding author)

* Jun-Yeong Kang and Yousung Yeon contributed equally to this work; Corresponding authors: Chang Choi

* This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00559546).

* This research was supported by the AI Computing Infrastructure Enhancement (GPU Rental Support) User Support Program funded by the Ministry of Science and ICT (MSIT), Republic of Korea (RQT-25-120226).

Received(February 20, 2026), Review Result(1st: March 7, 2026, 2nd: April 2, 2026), Accepted(June 13, 2026), Published(June 30, 2026)

modalities to accurately grasp the overall context and increase the precision of classification [1]. In the changing digital media environment, news is increasingly provided in a multimodal format that combines text with other media, such as images. Unimodal news has limitations in comprehending complex contexts and resolving ambiguous information within the text [2]. Multimodal AI analyzes data by fusing different forms of modality to overcome the limitations of a unimodal model. However, news articles often include images irrelevant to the text. Fusing these modalities without considering the semantic discrepancies between them causes the model to overfit to meaningless visual noise, ultimately degrading the final classification performance.

Multimodal AI news classification uses fusion techniques to determine the characteristics of the data. However, when text features used in fusion operations are encoded as plain text, it is difficult to preserve the main contextual information of the article until the final fusion process [3]. To address information loss in plaintext encoding, important contextual information can be preserved by applying named entity recognition (NER), which identifies unique entities in unstructured text, and keyword extraction techniques. However, simple concatenation provides no mechanism for dynamically adjusting weights based on the reliability of the input data; therefore, it cannot prevent predictions from being biased by visual noise that is independent of the text [4]. To mitigate visual noise bias, cross-attention mechanisms computing the similarity between individual text words and image regions have been proposed. However, even when meaningless noise images are input, cross-attention computes local similarities with the text for fusion. As a result, it has limitations in evaluating the overall reliability of the modality and dynamically adjusting the contribution weights.

In this paper, we propose an entity-aware bidirectional attention-based gated fusion that utilizes entity and keyword information to preserve text information and address the classification bias caused by visual noise. The proposed architecture consists of three stages: information extraction, bidirectional feature mapping, and dynamic gated fusion. The information extraction stage identifies unique entities and salient keywords within unstructured text using NER and keyword extraction techniques. This process establishes a reference point for evaluating the semantic relevance between modalities, ensuring that important context is not compromised by visual noise. Experimental results show that the proposed model improves the F1-score by up to 15.73% compared to the baseline.

The main contributions of this paper are as follows.

- We propose a multimodal fusion architecture that integrates the structured key information of news articles with a bidirectional cross-attention mechanism.
- We design a gated fusion mechanism that dynamically controls the relative contribution of each

modality to prevent model bias caused by visual noise.

- We introduce a preprocessing technique that preserves the key context of news articles by applying NER and keyword extraction to solve the information loss problem.

2. Related Work

Multimodal AI performs classification by learning correlations between different modalities and supplementing information that may be missing in unimodal models. However, when news texts are long and irrelevant images are included, simple concatenation of the two modalities is limited by semantic misalignment. For multimodal news classification, a multimodal framework must incorporate an interaction mechanism to address the structural limitations of data fusion methods, information loss in long texts, and visual noise issues. Related work is organized into three categories: multimodal fusion, context-preserving text representation, and bidirectional attention-based fusion mechanisms.

2.1 Multimodal fusion in news classification

Multimodal fusion methods are classified into early fusion, which concatenates extracted embeddings into a single vector, late fusion, which aggregates independent predictions from each modality, and hybrid fusion that maps inter-modal interactions within the architecture. Singhal et al. proposed an early fusion architecture that integrates visual and textual cues by concatenating feature vectors extracted from a pre-trained language model and an image classifier at a single layer [5]. Wang et al. introduced an auxiliary classifier to prevent the model from memorizing event-specific patterns and to ensure fused features remain unbiased by noise [6]. Nakamura et al. constructed a late fusion architecture that aggregates the final predicted scores from text and image models to validate its classification efficiency in a multi-class news setting [7].

Since early fusion concatenates extracted global features and late fusion merely aggregates predicted scores, both approaches fail to achieve semantic alignment between text and images [8]. Consequently, they cannot capture cross-modal interactions between words and objects within images, causing the final prediction to be biased by visual noise when generic photos unrelated to the news text are included [9]. To address the semantic misalignment and visual bias inherent in early and late fusion, this study proposes a hybrid fusion mechanism that dynamically learns inter-modal correlations. Gated fusion operates as a computational mechanism that implements the hybrid fusion architecture, dynamically determining the final contributions by learning the reliability of features extracted from each modality.

Therefore, this study adopts a hybrid fusion architecture combining bidirectional cross-attention and gated fusion to achieve semantic alignment while suppressing visual noise.

2.2 Entity and Keyword Extraction for Context Preservation

In multimodal news classification, text features are processed using a standard plain-text approach, which utilizes a pre-trained language model to encode the entire article into a global vector. However, when complex contexts and factual relationships are compressed into a single vector space, key information becomes dispersed due to irrelevant contextual noise, thereby diminishing its semantic weight. This leads to the loss of fine-grained semantics, posing a structural limitation that hinders precise mapping with image patches during the cross-attention interaction stage.

In multimodal news classification, various natural language processing (NLP) techniques have been explored to compress text context, aiming to prevent information loss in lengthy texts and facilitate alignment with visual information. Text summarization is utilized as a method to condense the salient content of an article into natural language sentences. Li et al. proposed a method for selecting key sentences by leveraging modality attention across text and image data [10]. However, as summarized text remains in a sequential format containing predicates and syntactic noise, it unnecessarily disperses attention weights during cross-attention for token-level alignment with image patches, thereby introducing cross-modal noise. Abstractive summarization reconstructs context through a probabilistic generation mechanism that predicts the next word based on statistical correlations. Consequently, it can induce hallucination, generating content absent from the original text, thereby compromising the factual integrity of the news article [11].

The effectiveness of identifying core issues and keywords from large-scale unstructured data through text mining has been validated in previous research [12][13]. Therefore, this study adopts NER and keyword extraction to preserve factual information and improve semantic alignment between text and images. By directly extracting noun and proper noun tokens from the source text, NER and keyword extraction preclude the possibility of information distortion associated with generative models and maintain the integrity of the source text. Providing only noise-filtered core semantic information optimizes the one-to-one mapping precision between visual objects and text tokens during the cross-attention stage. In conclusion, our method circumvents both the loss of fine-grained semantics inherent in plain-text encoding and the factual distortions caused by syntactic noise and hallucinations in conventional summarization techniques.

2.3 Bidirectional Cross-Attention and Gated Fusion Mechanisms

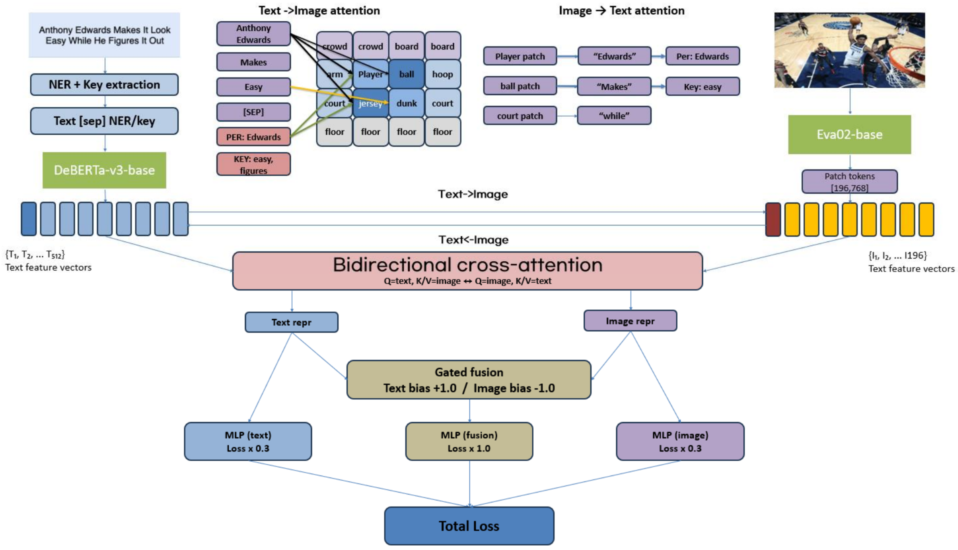
Hybrid fusion learns inter-modal interactions within the model to address the limitations of early fusion, which involves simple concatenation, and late fusion, which relies on independent predictions without such interactions. Unidirectional cross-attention utilizes a single modality as the query to select relevant features from the other modality. Jin et al. proposed att-RNN, which leverages news image features as a query to assign weights to important words in the text, thereby suppressing the impact of text noise [14]. However, since unidirectional cross-attention only reflects the feature vectors of the modality used as the query, it fails to synchronize the complex contexts of both modalities, leading to significant information loss.

Bidirectional cross-attention allows text and images to cross-reference each other to overcome the limitations of unidirectional interaction. Wu et al. enhanced news classification performance by employing a Multimodal Co-Attention Network (MCAN), which allows text and image features to reflect each other bidirectionally [15]. The softmax operation within attention forces weights to sum to 1.0 regardless of actual relevance, resulting in distortion where high weights are assigned to noise patches even when the semantic similarity between modalities is low. Qi et al. demonstrated that the inclusion of generic images irrelevant to the article body causes bidirectional cross-attention to assign high weights to non-informative visual patterns, thereby amplifying noise and degrading classification performance [16]. Bidirectional interaction architectures are limited in controlling model bias caused by irrelevant visual information because they cannot independently evaluate the reliability of each modality [17].

This study proposes a hybrid mechanism combining bidirectional cross-attention and gated fusion to resolve alignment errors and noise amplification. The bidirectional cross-attention in this study resolves the information loss of unidirectional methods by synchronizing and mutually complementing the contexts of both modalities [18]. Unlike the forced allocation used in softmax, the integrated gated fusion assigns weights by independently verifying the validity of each modality. By closing the gate for irrelevant image data, it blocks visual noise and implements semantic alignment based on the actual relevance between text and images.

3. Bidirectional Attention-Based Gated Fusion Architecture

This study proposes entity-aware bidirectional attention-based gated fusion to preserve the complex context of news data and control bias caused by visual noise.



[Fig. 1] Overview of the proposed architecture

[Fig. 1] illustrates the overall architecture of the proposed model. This architecture consists of a context-preserving preprocessing stage that extracts key entities and keywords to prevent context loss, a bidirectional cross-attention stage that performs alignment between two heterogeneous modalities, and a gated fusion stage that suppresses noise. The extracted named entities and keywords are combined with the original text and then encoded. During the gating operation, modality-specific bias initialization improves the initial training stability of the model and increases the reliability of the textual context by initializing the text and image biases [19].

Algorithm 1: Entity-aware Bidirectional Gated Fusion

Require: Raw text T_{raw}, Image I; Encoders f_{enc}, Classifiers f_{ds}

Ensure: Prediction $\hat{y}$, Total Loss L

1: Preprocessing: $T_{enh} \leftarrow$ Extract (Entities, Keywords) from T_{raw}

2: Encoding: X_{txt}, $X_{img} \leftarrow f_{DeBERTa}(T_{enh})$, f_{EVA02}(I)

3: Alignment:

  for each modality m $\in$ {txt, img} **do**

    $\hat{X}_m \leftarrow$ **Bi-Cross-Attention with residual connection**

  end for

4: Gating: Compute g_m with b_t = 1.0, b_i = -1.0

5: Fusion: $Z_{fused} \leftarrow \sum(g_m \odot \hat{X}_m)$

6: Optimization: Classify $\hat{y}$ & Calculate multi-head loss L

7: return $\hat{y}$, L

As summarized in Algorithm 1, the proposed entity-aware bidirectional gated fusion architecture outlines the overall processing pipeline from preprocessing to final optimization. The detailed mathematical formulations for each stage are described in the following subsections.

3.1 Context-Preserved Feature Extraction and Embedding

The context-preserving preprocessing stage extracts the core context of the text from the input data, including headlines, captions, summaries, and article bodies. NER and keyword tokens are concatenated with the original text to reinforce salient information, effectively mitigating the contextual sparsity and limited semantic representation inherent in short text fields [20][21]. The salient information, combined with the original text at a token level, induces the model's attention mechanism to assign high weights to key entities and keywords, thereby compensating for the limited semantic representational power of short texts.

Following the preprocessing stage that reinforces the salient information of the text by combining NER and keywords, the image and text modalities are transformed into token-level representations of the same dimension through encoders. The text encoder utilizes DeBERTa-v3-base [22][23], which learns complex word dependencies through a disentangled attention mechanism that computes word content and relative position information separately.

The image encoder uses EVA02, which is pre-trained by randomly masking portions of an image and then reconstructing the masked regions by inferring them based on visual context [24]. The text and image encoders used in this study produce feature vectors with the same dimensionality, allowing the two heterogeneous modalities to be directly integrated without conversion loss or the overhead of additional parameter computation. The text tokens x_{txt} and image tokens x_{img} generated as in Equation [1] and Equation [2] have sequence lengths of L_t and L_i respectively, and are used as inputs for bidirectional interaction.

$$X_{txt} = f_{DeBERTa}(Text_{NER,Key,Raw}) \in \mathbb{R}^{\mathbb{B} \times L_{t \times 768}} \qquad \text{Equation [1]}$$

$$X_{img} = f_{EVA02}(Image) \in \mathbb{R}^{\mathbb{B} \times L_i \times 768} \qquad \text{Equation [2]}$$

3.2 Bidirectional Cross-Attention Mechanisms

The feature vectors x_{txt} and x_{img} extracted from the individual encoders are fed into the bidirectional cross-attention module, allowing text and image to simultaneously cross-reference each other's features to learn inter-modal interactions.

As shown in Equation [3] and Equation [4], in text-to-image attention, the text is used as the query and the image as the key and value to extract visual features from the perspective of the text. The NER and keyword tokens injected into the query serve as a reference when performing the dot-product operation with the key patches of the image. Consequently, similarity scores for background patches with low visual relevance are suppressed, while high attention weights are assigned to key object patches that visually align with the text entities. This yields the noise-refined text tokens $\hat{x}_{txt}$.

As shown in Equation [5] and Equation [6], in image-to-text attention, the image is used as the query, and the text as the key and value to extract text features $\hat{x}_{img}$ from the perspective of the image. $W_{Q,}W_{K,}W_V$ are learnable weight matrices, and $\sqrt{d}$ is a scaling factor that normalizes the magnitude of the dot product to ensure a stable softmax distribution.

$$Q_t = X_{txt}W_Q^t,\ K_i = X_{img}W_K^i,\ V_i = X_{img}W_V^i \quad \text{Equation [3]}$$

$$\hat{X}_{txt} = softmax(\frac{Q_t K_i^T}{\sqrt{d}})V_i \quad \text{Equation [4]}$$

$$Q_i = X_{img}W_Q^i,\ K_t = X_{txt}W_K^t,\ V_t = X_{txt}W_V^t \quad \text{Equation [5]}$$

$$\hat{X}_{img} = softmax(\frac{Q_i K_t^T}{\sqrt{d}})V_t \quad \text{Equation [6]}$$

3.3 Gated Fusion with Modality-Specific Bias

The two feature vectors $\hat{x}_{txt}$, $\hat{x}_{img}$ synchronized through bidirectional cross-attention are fed into the gated fusion module to control noise amplification caused by the zero-sum nature of the softmax operation. The gating mechanism proposed in this study independently verifies the validity of each modality to assign weights [19][25]. In news data, named entities within the text exhibit a strong correlation with the classification labels. However, there are exceptions where textual information is ambiguous or visual cues play a pivotal role. This study applies a modality-specific bias initialization technique to prevent static bias toward a specific modality. Although it initially reflects the statistical trends of the data, the model updates its weights and biases through backpropagation during the training process to adjust the activation level of each gate. As shown in Equation [7] and Equation [8], the bias b_t of the text gate g_{txt} is initialized to +1.0, and the bias g_{img} of the image gate b_i is initialized to − 1.0.

$$g_{txt} = \sigma(W_{gt}[\hat{X}_{txt};\hat{X}_{img}] + b_t) \quad \text{Equation [7]}$$

$$g_{img} = \sigma(W_{gi}[\hat{X}_{txt};\hat{X}_{img}] + b_i) \quad \text{Equation [8]}$$

This asymmetric bias initialization maximizes the contribution of highly reliable textual information in the early stages of training, while minimizing the incorporation of noise from irrelevant image patches. Finally, as shown in Equation [9], the calculated independent gate values are multiplied element-wise ⊙ with their respective feature vectors to generate the fusion vector z_{fused}. The generated z_{fused} is then passed to the classifier layer to determine the final news category [26].

$$Z_{fused} = g_{txt} \odot \hat{x}_{txt} + g_{img} \odot \hat{x}_{img} \quad \text{Equation [9]}$$

3.4 Objective Function and Optimization

This study applies an auxiliary loss function that prevents the modality collapse phenomenon and forces each encoder to independently extract visual and textual contexts. In addition to the final fusion vector z_{fused} obtained through gated fusion, the individual feature vectors $\hat{x}_{txt}$ and $\hat{x}_{img}$ derived from cross-attention pass through a multi-layer perceptron classifier to calculate prediction losses [9].

The total loss, L_{Total}, is defined as a weighted sum based primarily on the fusion loss L_{fusion}, incorporating independent classification losses L_{txt} and L_{img} for text and image with a weight factor λ, as shown in Equation [10]. In this study, the loss weights λ_{txt} and λ_{img} for each modality are set to 0.3 based on the validation dataset to perform normalization.

$$L_{Total} = L_{fusion} + \lambda_{txt} L_{txt} + \lambda_{img} L_{img} \quad \text{Equation [10]}$$

4. Experimental Results

This section evaluates the news category classification performance of the proposed entity-aware bidirectional attention-based gated fusion in a multimodal news data environment. The experiments verify whether the proposed model maintains robust classification performance in environments where semantic discrepancies between text and image modalities or visual noise exist. This study compares the performance of the proposed architecture against the baselines presented in the original N24News dataset paper to verify its effectiveness.

4.1 Experiment Setup and Dataset

This study utilizes the N24News dataset, a multimodal news dataset, to verify the generalized news classification performance of the proposed model. The N24News dataset comprises headlines, captions, abstracts, and body texts from New York Times news articles along with their associated images, and is categorized into 24 classes. The experimental framework consists of nine distinct configurations categorized into text-only, image-only, and multimodal settings. Unimodal experiments are conducted to compare the performance differences between backbone models when processing each modality independently. This experiment utilizes a total of 61,235 data samples, which are partitioned into training, validation, and test sets of 48,988, 6,123, and 6,124 samples, respectively, according to an 8:1:1 ratio.

We utilize DeBERTa-v3 and EVA02 as the text and image backbones, respectively. All experiments are conducted on a system equipped with an Intel Xeon® w5-2455X processor and an NVIDIA GeForce RTX 4090 24 GB GPU. The implementation was developed using Python 3.10 and the PyTorch framework. The learning rate for model training is set to 2e-5, and all experiments are conducted for a maximum of 15 epochs. Early stopping is applied if the validation loss fails to improve for five consecutive epochs. The final performance is evaluated during the testing phase using the model weights that achieved the highest validation accuracy. Accuracy, defined in Equation [11], and the F1-score, defined in Equation [12], are used as evaluation metrics to objectively assess model performance in the multi-class classification task.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \text{Equation [11]}$$

$$F1-score = \frac{2\times TP}{2\times TP+FP+FN} \qquad \text{Equation [12]}$$

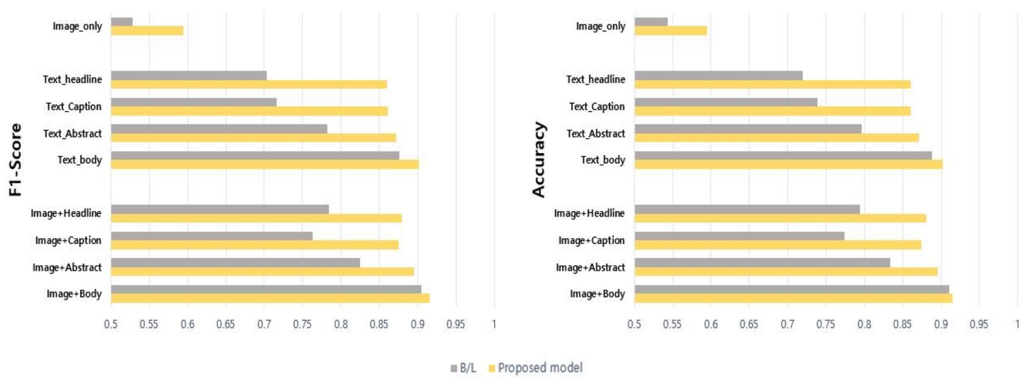
4.2 Comparison of Proposed Method with Baselines

This study compares the performance of the proposed architecture against the baselines presented in the original N24News paper to verify its effectiveness. The existing baselines utilize CNN-based ResNet and RoBERTa with absolute positional encoding as feature extractors [27][28]. ResNet is limited in its ability to capture the global visual context of an entire image because it sequentially processes only limited, narrow regions using small filters. RoBERTa faces limitations in capturing intricate semantic relationships between words in long documents, such as news articles, because it only learns fixed positional information for each word.

This study replaces the feature extractors with DeBERTa-v3 and EVA02 to overcome the limitations of ResNet and RoBERTa. DeBERTa-v3 precisely identifies the relationships between key entities in news articles by separately calculating word content and relative positions. EVA02 applies self-attention to calculate the relationships between multiple image patches simultaneously, extracting the global visual context that the baseline ResNet has difficulty capturing.

[Table 1] Performance comparison between the Baseline and the Proposed model. The values in parentheses (+) indicate the performance improvement over the baseline model

Feature extractor		Baseline		Proposed model	
Modality	**Input data**	**F1-score**	**Accuracy**	**F1-score**	**Accuracy**
Image	Image_only	0.5280	0.5434	0.5929 (+0.0649)	0.5944 (+0.0510)
Text	Text_Headline	0.7031	0.7198	0.8604 (+0.1573)	0.8602 (+0.1404)
	Text_Caption	0.7156	0.7387	0.8608 (+0.1452)	0.8610 (+0.1223)
	Text_Abstract	0.7819	0.7965	0.8718 (+0.0899)	0.8715 (+0.0750)
	Text_Body	0.8765	0.8886	0.9020 (+0.0255)	0.9017 (+0.0131)
Multimodal	Image+Headline	0.7842	0.7941	0.8798 (+0.0955)	0.8803 (+0.0862)
	Image+Caption	0.7633	0.7745	0.8753 (+0.1120)	0.8748 (+0.1003)
	Image+Abstract	0.8252	0.8333	0.8956 (+0.0704)	0.8955 (+0.0622)
	Image+Body	0.9044	0.9108	0.9152 (+0.0108)	0.9154 (+0.0046)



[Fig. 2] Comparison of news classification performance according to different input modalities

The proposed model demonstrated consistent performance improvements over the baseline across all experimental settings. As shown in [Table 1] and [Fig. 2], the proposed model outperformed the baseline across all nine input conditions, including both unimodal and multimodal settings. Specifically, for the headline input, the model achieved a significant performance increase of 14.04% in accuracy and 15.73% in F1-score.

5. Conclusion

Online news data is presented in a multimodal format that integrates diverse media, such as images, alongside textual content. However, visual noise irrelevant to the article context or ambiguous text can hinder accurate category classification. Concatenation-based fusion methods suffer from limitations, including information degradation during modality integration and the incorporation of visual noise from heterogeneous data into the learning process. This paper proposes an entity-aware bidirectional attention-based gated fusion architecture to overcome the limitations of multimodal fusion. The proposed model performs bidirectional cross-attention based on named entities and keywords extracted during the preprocessing stage to refine visual features and suppress noise. The gated fusion stage resolves the optimization imbalance between modalities and ensures training stability by applying bias initialization and auxiliary loss functions. Experimental results using the N24News dataset demonstrate that the proposed model achieved performance gains of 14.04% in accuracy and 15.73% in F1-score over the existing baseline, confirming that it achieves stable news classification performance in environments characterized by semantic inconsistency and noise. The proposed architecture is expected to be applicable in various intelligent media applications, such as fake news detection and personalized news recommendation services.

References

[1] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, Feb. 2018, doi: 10.48550/arXiv.1705.09406.

[2] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, "BreakingNews: Article annotation by image and text processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2410-2424, Oct. 2018, doi: 10.1109/TPAMI.2017.2721945.

[3] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373-63394, May 2019, doi: 10.1109/ACCESS.2019.2916887.

[4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, Sep. 7-11, 2017, pp. 1103-1114, doi: 10.18653/v1/D17-1115.

[5] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumar, and K. Yamanishi, "SpotFake: A multi-modal framework for fake news detection," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Singapore, Sep. 11-13, 2019, pp. 39-47, doi: 10.1109/BigMM.2019.00-44.

[6] Y. Wang et al., "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*, London, UK, Aug. 19-23, 2018, pp. 849-857, doi: 10.1145/3219819.3219903.

[7] K. Nakamura, S. Levy, and W. Y. Wang, "r/Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection," in *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, May 11-16, 2020, pp. 6149-6158.

[8] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, Seattle, WA, USA, Oct. 12-16, 2020, pp. 1122-1131, doi: 10.1145/3394171.3413678.

[9] W. Wang, R. Tran, and N. Bertels, "What makes training multi-modal classification networks hard?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, Jun. 14-19, 2020, pp. 12695-12705, doi: 10.1109/CVPR42600.2020.01271.

[10] H. Li, J. Zhu, J. Zhang, X. He, and C. Zong, "Multimodal sentence summarization via multimodal selective encoding," in *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain, Dec. 8-13, 2020, doi: 10.18653/v1/2020.coling-main.496.

[11] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1-38, Mar. 2023, doi: 10.1145/3571730.

[12] R. Mihalcea and P. Tarau, "TextRank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, Jul. 25-26, 2004, pp. 404-411.

[13] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, San Diego, CA, USA, Jun. 12-17, 2016, pp. 260-270.

[14] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM International Conference on Multimedia*, Mountain View, CA, USA, Oct. 23-27, 2017, pp. 795-816, doi: 10.1145/3123266.3123454.

[15] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online, Aug. 1-6, 2021, pp. 2560-2569, doi: 10.18653/v1/2021.findings-acl.226.

[16] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE International Conference on Data Mining (ICDM)*, Beijing, China, Nov. 8-11, 2019, pp. 518-527, doi: 10.1109/ICDM.2019.00062.

[17] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and

sentence matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, Jun. 14-19, 2020, pp. 10941-10950.

[18] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, Nov. 3-7, 2019, pp. 5100-5111, doi: 10.18653/v1/D19-1514.

[19] J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated multimodal units for information fusion," in *Proceedings of the 5th International Conference on Learning Representations Workshop*, Toulon, France, Apr. 24-26, 2017, doi: 10.48550/arXiv.1702.01992.

[20] Z. Wang, X. Shan, X. Zhang, and J. Yang, "N24News: A new dataset for multimodal news classification," in *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, Marseille, France, Jun. 20-25, 2022, pp. 6982-6990.

[21] K. D. Park, S. D. Ju, and M. C. Hwang, "A Study on Improvement of Integrated Culture Voucher Management through Big Data Analysis," *Journal of Digital Media & Culture Technology*, vol. 9, no. 2, pp. 201-214, Jun. 2022, doi: 10.29056/jdaem.2022.06.08.

[22] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," in *International Conference on Learning Representations (ICLR)*, Virtual, Apr. 25-29, 2022, doi: 10.48550/arXiv.2111.09543.

[23] W. S. Lee, "Performance Evaluation of Technical Patent Classification Using BERT," *Journal of Digital Media & Culture Technology*, vol. 13, no. 2, pp. 277-285, Apr. 2024, doi: 10.29056/jncist.2024.04.12.

[24] Y. Fang et al., "EVA-02: A visual representation for neon genesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, Jun. 18-22, 2023, doi: 10.1016/j.imavis.2024.105171.

[25] W. Y. Shin and Y. H. Kim, "Improvement of Art Distribution Structure using Blockchain Technology," *Journal of Digital Media & Culture Technology*, vol. 11, no. 3, pp. 253-262, Jun. 2022, doi: 10.29056/jncist.2022.06.03.

[26] W. S. Lee, "Portfolio Study based on Deep Deterministic Policy Gradient Model," *Journal of Digital Media & Culture Technology*, vol. 11, no. 3, pp. 287-298, Jun. 2022, doi: 10.29056/jncist.2022.06.06.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 27-30, 2016, pp. 770-778.

[28] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China, Nov. 3-7, 2019, pp. 1-13, doi: 10.48550/arXiv.1907.11692.