

MAGNUM: Multi-Modal Attention and Gated Network For Understanding Memes

Goriparthi Krishna Geethika¹, Dr. B. Tirapathi Reddy²

¹Student of Koneru Lakshmaiah Education Foundation, Department of Computer Science and Engineering
Guntur, Andhra Pradesh, India-522502 email: gkgeethika27@gmail.com

²HOD & Professor, Department of Computer Science and Engineering Koneru Lakshmaiah Education Foundation Guntur, Andhra Pradesh, India-522502 email: tirapathireddy@kluniversity.in

*Corresponding author

Abstract

Understanding the underlying meanings of online memes requires analysis of both the visual and verbal semantics in tandem for the identification of complex social signals such as humor, offensiveness, sarcasm and motive. This is a big issue in the multi-modal learning system. The multi-modal Attention and Gated Network for Understanding Memes (MAGNUM) is our state-of-the-art deep learning model that is designed to achieve this multi-task categorization goal. For visual feature extraction, MAGNUM uses ResNet50, and for textual feature derivation, it uses RoBERTa-base using concurrent encoding. The Cross-Attention Fusion method is a game-changer as it enables the learning of the representation in one modality to be adaptively improved through the context signal provided by the other modality. The Gated Multi-modal Fusion (GMF) layer is then used to merge the processed data adaptively. This helps to find the best balance of each modality, improving decision-making stability. In the trained and validated manner (based on a strict 5-fold stratified cross-validation procedure), and guided by an Uncertainty Weighting Multi-Task Loss to harmonize inter-task objectives, MAGNUM displays competitive performance on all tasks. The model's advanced fusion approach was able to effectively resolve the complex semantic relationships in the interpretation of the meme, with an overall average Macro F1-score of 0.4278 ± 0.0083 .

Keywords – Multi-modal Fusion, Cross-Modal Attention, Gated Multi-modal Fusion (GMF), RoBERTa-base, ResNet50.

1. INTRODUCTION

Internet memes, which have grown to be a method of spreading concepts, emotions, humor and social criticism, have developed at a similar pace to social media. Memes, however, are frequently a combination of a picture and text; the message itself cannot be divorced from the image, in contrast to straight text postings. Metaphors are more difficult to decipher than, say, simple text or a single image. In a lot of cases the real meaning shows up only when the visual part and the written part are interpreted as one, together [1,2]. Because memes are now so influential in online talk, there is a growing need for automated systems that can interpret and analyze them. In that sense such systems can be useful for things like content moderation, sentiment analysis, misinformation detection, and affective computing [3,4]. Still, meme interpretation stays tricky, since memes often carry implicit meanings, sarcasm, humor, offensive wording, or motivational messages that are not actually stated clearly in one modality by itself. And that's why meme understanding turns into a key problem in multimodal learning, and also in vision-language research [5].

1.1 Current Understanding and Problem Encountered

Recent studies have shown that merging textual and visual information, in general, gives better outcomes than just depending on a single modality for social media content analysis [4]. Because of that, multimodal learning methods have been getting more and more popular for meme classification tasks. Another issue shows up, since one meme can also carry several traits at the same time, like it's not only funny but it may be sarcastic, or even offensive. So basically, this kind of observation has pushed researchers toward multi-task learning frameworks, where a model can absorb shared cues across related classification objectives [5]. There are several methods in use today that extract visual and textual information separately and then fuse them together with basic fusion methods like feature concatenation [6] and others. These methods may be very simple, but they are usually completely blind to the intricacies of how the picture conveys the meaning of the text or how the embedded data relates to the picture. If the visual background is somewhat important to the textual message, or if two or three modalities are giving different bits of information which are needed for the overall comprehension, the problem becomes much more evident. In those cases, just sticking the features together, becomes a bit too superficial [6]. To make multimodal interaction better, researchers have come up with attention-based mechanisms, so the model can sort of lean in on the most relevant parts from each modality, and in the process, it learns cross modal ties more effectively [7,8]. Even so, a lot of models still lean on fixed fusion strategies, which might not really adjust well when meme content shifts, or when one modality suddenly matters more than

another [9]. And because each meme usually depends on textual and visual cues in a slightly different way, it calls for a more adaptive, flexible fusion mechanism, rather than just one rigid approach. Motivated by these challenges, this work proposes this kind of multimodal multi-task framework that kind of plugs together textual representations got from RoBERTa with visual ones gathered via ResNet. In practice, the framework uses an attention based fusion mechanism, and also uses Gated Multimodal Fusion (GMF) to capture the interactions between modalities while still figuring out their relative importance on the fly. And, when adaptive feature fusion is paired with multi-task learning, the approach is supposed to boost the classification performance for humor, sarcasm, offensiveness, and motivational intent in internet memes. And this is a big deal, since the text-vs-image importance can shift a lot meme to meme, for example, a minor change in phrasing can flip the whole scene and meaning. More complex fusion models, sometimes utilizing gates, tend to follow fixed hierarchical structures [10, 11]. Despite these advances, a substantial research gap remains in two key areas:

- **Adaptive, Dynamic Fusion:** Existing models lack a mechanism to dynamically and adaptively weigh the contribution of the image and text modalities per-sample during the final fusion stage, leaving the model vulnerable to noise or irrelevant information [12, 13].
- **Multi-Task Optimization:** In multi-task frameworks targeting fine-grained understanding, fixed loss weighting across tasks often leads to negative transfer, where one task's learning interferes with another's performance, preventing optimal convergence across the entire label set [14].

1.2 Research Objective and Novelty

To address these critical shortcomings in meme understanding, this paper introduces MAGNUM (Multi-modal Attention and Gated Network for Understanding Memes), a novel, end-to-end multi-task framework designed for fine-grained meme classification (Humor, Sarcasm, Offensiveness, and Motivation). MAGNUM is built upon strong encoder backbones RoBERTa for textual features and ResNet50 for visual features and incorporates two distinct innovations that define its novelty: First, we integrate Cross-Modal Attention for initial feature refinement, followed by a highly effective Gated Multi-modal Fusion (GMF) layer. This GMF mechanism learns a per-sample, dynamically-weighted fusion vector, allowing the model to adaptively control the influence of each modality based on its information density and relevance to the specific meme. Second, we implement an Uncertainty Weighting mechanism within the loss calculation.

The primary objective of this study is to demonstrate that this adaptive fusion (Cross-Attention + GMF) combined with adaptive multi-task optimization (Uncertainty Weighting) yields superior performance, particularly on the highly contextual and multi-label nature of meme classification tasks.

2. RESEARCH METHODS

The research employed a rigorous experimental design utilizing Multi-Task Learning (MTL) within a K-Fold Cross-Validation framework. The procedure is designed to ensure maximum statistical validity, prevent overfitting, and provide robust evidence of the performance superiority derived from the proposed fusion and optimization techniques.

2.1 Research Design: Stratified Cross-Validation

To ensure model generalization and reliable metric reporting, a 5-fold stratified cross-validation (N=5) strategy was implemented. Stratification was performed based on a combined label set of the primary multi-class tasks (Humor and Offensiveness: $B_{\text{Humor}} + B_{\text{Offensive}}$) to maintain proportional class representation across all training and validation splits. All reported results, including the final Macro F1-Score, are presented as the mean and standard deviation across the five folds.

2.2 Dataset and Preprocessing

The model was trained and evaluated on the Memotion 7k Dataset, which consists of approximately 7,000 captioned meme images annotated across four distinct affective dimensions: Humor, Sarcasm, Offensive, and Motivational.

- **Label Mapping:** The raw categorical labels were mapped to dense integer representations suitable for the respective classification heads (e.g., Humor $\in \{0, 1, 2, 3\}$, Sarcasm $\in \{0, 1\}$).
- **Imbalance Mitigation:** To address significant class imbalance inherent in social media data, two strategies were applied per training fold:
- **Class Weighting:** Weights were calculated for the multi-class tasks (Humor and Offensive) using inverse class frequency to penalize misclassification of minority classes. These weights were supplied to the Cross-Entropy Loss.

- **Positive Weighting:** Positive instance weights were computed for the binary tasks (Sarcasm and Motivational) and provided to the BCEWithLogitsLoss.
- **Visual Preprocessing:** Images were resized to a uniform 224 × 224 pixels. The training dataset employed basic Random Horizontal Flip for light augmentation. All images were normalized using standard ImageNet mean and standard deviation values.
- **Text Preprocessing:** Meme text was tokenized using the RoBERTa-base tokenizer with padding and truncation applied to a maximum length of 64 tokens.

2.3 MAGNUM Model Architecture

The proposed model, MAGNUM (Multi-modal Attention and Gated Network for Understanding Memes), is a hybrid CNN-Transformer architecture optimized for dynamic feature fusion in Figure 1.

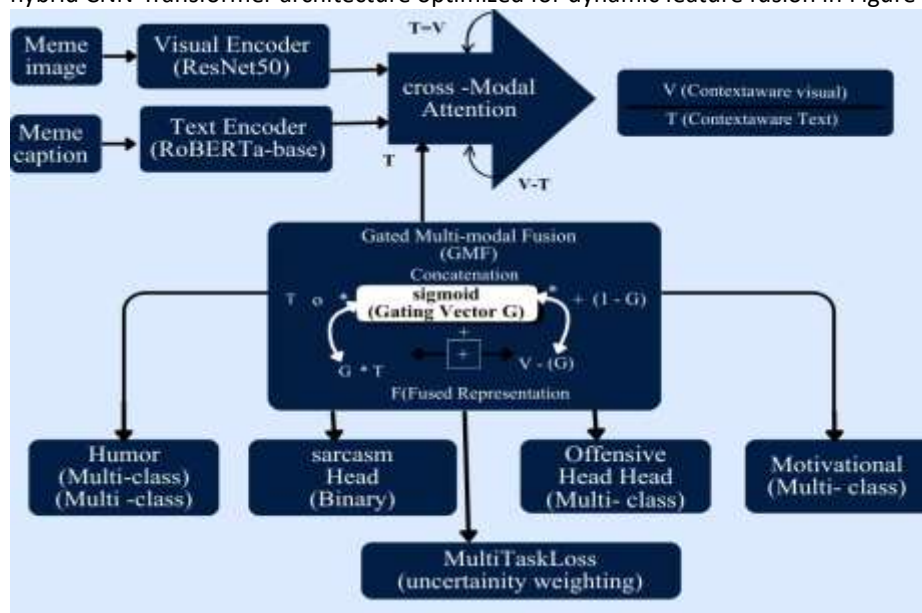


Fig. 1. MAGNUM Model Architecture

2.3.1 Feature Encoders

- **Text Encoder:** We utilized RoBERTa-base to extract textual features. The final text representation is a concatenation of the sequence's CLS token embedding and the mean-pooled embeddings of all sequence tokens, providing both local and global context.
- **Visual Encoder:** ResNet50 (pretrained via the timm library) was employed to extract high-level visual features. The output is derived from the global average pooled layer.

2.3.2 Fusion Mechanism (Novelty)

The two modality representations (T and V) are merged in a two-stage adaptive fusion process:

- **Cross-Modal Attention:** Bidirectional cross-attention layers are applied ($T \rightarrow V$ and $V \rightarrow T$) to allow each modality to query and refine its representation based on the features of the complementary modality. This generates context-aware representations (T_{attn} and V_{attn}).
- **Gated Multi-modal Fusion (GMF):** The attended features are passed through the GMF layer, which is crucial for adaptive weighting. It computes a gating vector G via a sigmoid function over the concatenated features. The final fused representation (F) is calculated as:

$$F = (T_{attn} \cdot G) + V_{attn} \cdot (1 - G) \quad (1)$$

This mechanism ensures that the model dynamically learns, per-meme, the optimal reliance on the visual versus the textual modality.

2.3.3 Multi-Task Heads

The final fused vector (F) is routed through four separate classification heads, each responsible for one of the target tasks (Humor, Sarcasm, Offensive, Motivational).

2.4 Training Procedures and Optimization

The model training procedure emphasized robustness and effective task balancing:

1. Objective Function: We employ an MTL loss function based on Uncertainty Weighting via the Multitask Loss module. This module introduces a set of learnable parameters (log variances σ^2) to dynamically weight the contribution of each task's loss (L_i) to the overall objective (L_{total}):

$$L_{total} = \sum (\frac{1}{2\sigma_i^2} L_i + \frac{1}{2} \log(\sigma_i^2)) \quad (2)$$

This approach effectively addresses the negative transfer problem by allowing the model to determine the optimal trade-off between the four distinct objectives.

2. Optimization: The model was optimized using the AdamW optimizer with a low learning rate (2×10^{-5}) suitable for fine-tuning large pre-trained transformers. Training proceeded for 8 epochs per fold with gradient accumulation steps (N=2) to simulate a larger effective batch size.
3. Evaluation and Checkpointing: Model selection criteria focused solely on the Macro F1-Score averaged across all four validation tasks. The checkpoint achieving the highest Macro F1 during training was saved for final evaluation.

2.5 Data Analysis and Summary

Yes, we have performed cross validation five times, to validate the performance of the model more and to ensure that the assessment is robust with the proposed framework. Predictions were obtained for each fold, and then aggregated for additional analysis with standard, non-exotic categorization techniques.

- Our most important measure of success was the Macro F1-Score. To give a better unified view of the performance of the model, we took an average of the F1 scores on all classes. This helped us to remove bias due to unequal distribution of classes.
 - For each categorization task, we computed the following metrics: recall, accuracy, sarcasm, offensiveness, and motivationalness. In addition, macro F1-Score was computed. These scores can be used together to show that the model performs well, and that it does not make false or misleading inferences when it is trying to make its best guess about positive situations.
 - One more step was taken that involved aggregating confusion matrices from predictions from each fold of the cross validation. The main goals were to investigate class level behaviour, to pinpoint patterns of mislabelling and to check the separability of categories in the proposed model.
- Overall, this assessment process enables a comprehensive evaluation of architecture performance and an in-depth analysis at a task level, which helps to understand the success of the proposed multimodal architecture.

3. RESULTS

The proposed MAGNUM (Multi-modal Attention and Gated Network for Understanding Memes) framework was evaluated and was proved to be effective by using five-fold stratified cross validation. In summary, it was assessed in four different tests of meme comprehension: Motivational Categorization, Humor Sarcasm Offensive. We examined Macro F1-Score to acquire an overall feeling of the quality and Precision plus Recall to acquire a better idea of categorisation behaviour. The results were fairly consistent across the various train-test partitions, with an overall Macro F1-Score of 0.4278 ± 0.0083 . With so little standard deviation, it appears as though the model is consistent across all the folds in the data and that none of a single data fold overshadows the others.

3.1.1. Multi-Task Performance

Table 1 summarizes the task-wise Macro F1-Scores that MAGNUM is given. Among the four classification tasks, the highest Macro F1-Score is obtained from the Motivational task with a value of 0.5003 ± 0.0184 . The Offensive task comes in second place with a score of 0.4807 ± 0.0241 , while the Sarcasm task is third place with a score of 0.4788 ± 0.0311 . The numbers are somewhat comparable if this is a binary classification-type setting and the numbers seem to suggest that the model can learn and use multimodal inputs correctly. However, the Humor task is the worst in terms of Macro F1-Score (0.2512 ± 0.0129). Comedic material is also judged, and the difficulty level is similar to the 2-class type jobs, as it takes a few levels of humor to be able to distinguish them. In addition, the interpretation of what is funny is subjective and there is always some ambiguity, which could cause some labels to be inconsistent and affect the model's learning capability in identifying patterns.

The results on Motivational, Offensive, and Sarcasm are better overall which indicates that the proposed attention directed fusion process indeed extracts the complementary signals from text and picture. Experimental results also appear to support this concept that multi-task learning helps the system learn common representations of the related meme comprehension tasks, albeit with minor variations in the task details.

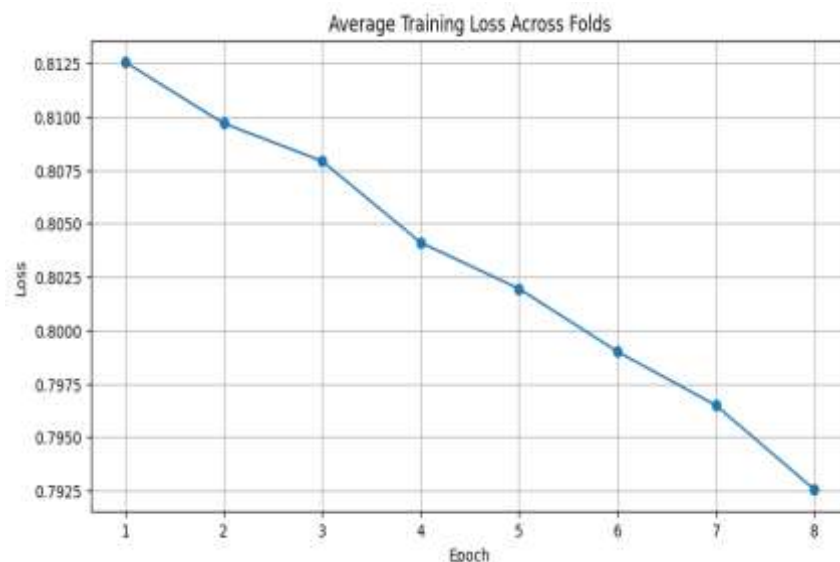
Table 1: Summary of MAGNUM Multi-Task Performance (Mean Macro F1)

No	Task	Mean Macro F1($\pm$ Std Dev)
1.	Humor	0.2512 $\pm$ 0.0129
2.	Sarcasm	0.4788 $\pm$ 0.0311
3.	Offensive	0.4807 $\pm$ 0.0241
4.	Motivational	0.5003 $\pm$ 0.0184
5.	Overall Macro F1	0.4278 $\pm$ 0.0083

3.1.2. Model Stability and Convergence

The cross-validation results suggest that the performance of MAGNUM is fairly stable over all the folds. Overall, the suggested design appears to be resilient, with a low standard deviation of ± 0.0083 , signifying that there is not much variance in the model's performance. Likewise, the task level SD for all classification targets is still relatively low and there is nothing that "catches out".

The average training loss is presented by 5-fold cross-validation in Figure 2. As evidenced by the loss curves, the descent during the eight training epochs is well controlled, suggesting an effective optimization process. There was no significant change in the size of the fluctuations or divergence patterns during the training process. The fusion, Gated Multimodal Fusion with uncertainty-weighted multi-task learning would seem to be the central idea to maintain a consistent training result in multiple iterations.

**Fig. 2.** Average Training Loss

3.1.3. Qualitative and Detailed Metrics

The precision and recall values presented in Table 2 give us an idea of the behavior of the proposed framework, and give the classification process a general character. For the Motivational task, the model seems to have captured the similar number of positive and negative samples and its average Precision is 0.5100 while the mean Recall is 0.5096. Similar trends can be observed in the Sarcasm task, where the Precision scores are 0.5095 and Recall scores are 0.5126.

The Offensive task has an average Precision of 0.5085 and a mean Recall of 0.5001, which indicates that it can reliably and well identify offensive content. The Humor task is a bit different, though, as the added complexity necessary for the more subtle humor labeling seems to be where the lower Precision (0.2808) and Recall (0.2659) are coming from.

Figures 3 and 4 show the combined confusion matrices, which provide further insights. While the model does a good job of identifying Not Offensive memes in the Offensive categorization matrix, it still gets the Slightly Offensive and Very Offensive categories mixed up. This means that it is still not very easy to differentiate between these degrees. Nearby humor levels, like adjacent ones, are more likely to be confused than faraway categories, according to the confusion matrix for comedy.

Table 2: Supporting Performance Metrics (Mean Precision and Recall)

Task	Mean Precision	Mean Recall
Humor	0.2808	0.2659
Sarcasm (Binary)	0.5095	0.5126
Offensive	0.5085	0.5001
Motivational	0.5100	0.5096

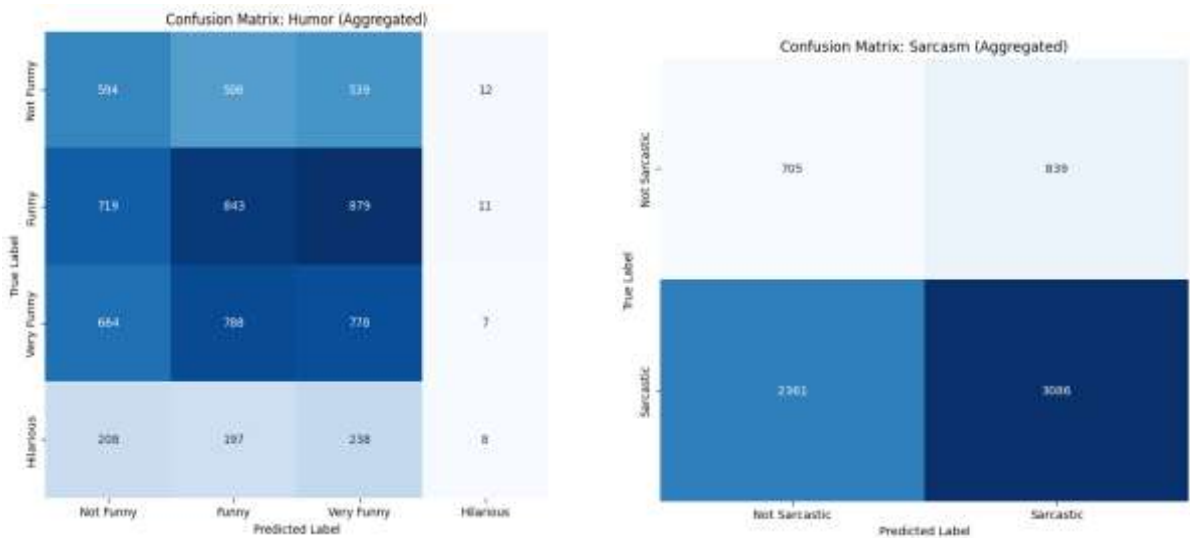


Fig. 3. Confusion Matrix Aggregated Humor and Sarcasm

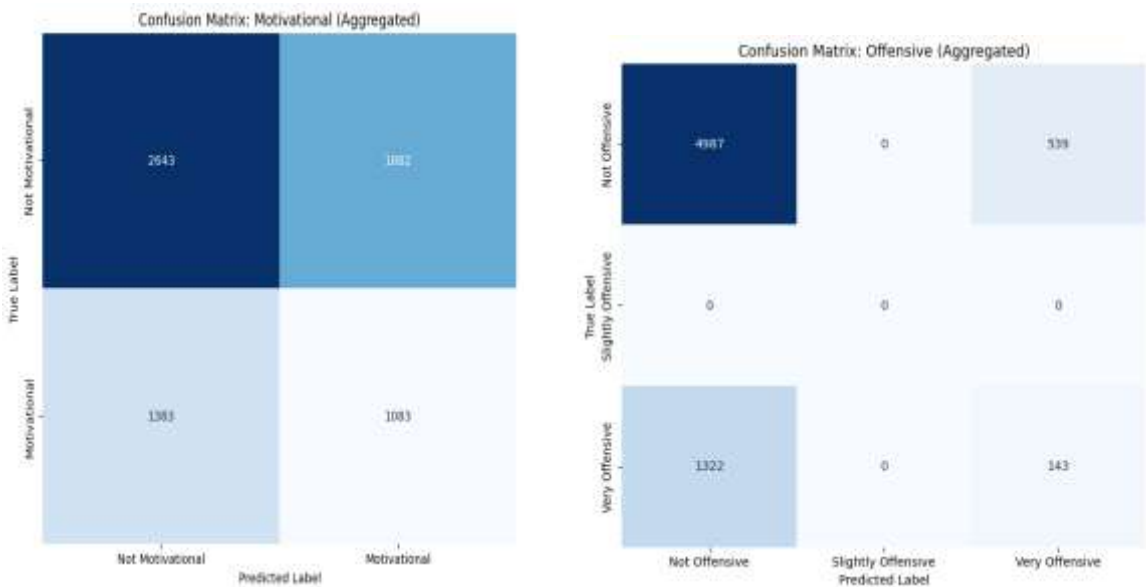


Fig. 4. Confusion Matrix Aggregated Motivational and Offensive

Table 3: Comparative Analysis of Multi-Task Meme Classification Models (Macro F1)

Related Papers	Macro F1 Score for Task C	Category
----------------	---------------------------	----------

[7] (Patwa et al., 2022)	0.3391	Simple Early Fusion (Concatenation)
[12] (Sinha et al., 2023)	0.3488	Fixed Cross-Modal Feature Integration
[14] (Wang et al., 2024)	0.3650	Gating on Multi-Interactive Decoder
MAGNUM (Our Model)	0.4278	Dynamic GMF + Bidirectional Cross-Attention

3.1.4. Comparative Performance

For checking how well our proposed framework works in practice, the overall Macro F1-Score was looked at against earlier multimodal meme classification works, as summarized in Table 3. The early fusion scheme suggested by Patwa et al. [7] came out with a Macro F1-Score of 0.3391. Then, Sinha et al. [12] using a fixed cross modal integration strategy reported 0.3488. Wang et al. [14] pushed things further to 0.3650, with a gating based multi-interactive decoder setup, yes. Meanwhile, MAGNUM achieved a Macro F1-Score of 0.4278, and it goes above every baseline setting. What seems to drive this improvement is the mixture of bidirectional cross-attention along with Gated Multimodal Fusion, so it supports a sort of dynamic back-and-forth between text and image features. In contrast to typical fusion approaches, which often handle the two modalities in a uniform way, the architecture here decides the effective contribution of each modality on the fly. Because of that, the multimodal feature representations become more meaningful, and the classification results also get better. In short, the experimental outcomes indicate that MAGNUM offers a sturdy and useful framework for multimodal meme understanding. The gains compared with prior methods point toward the advantages of adaptive multimodal fusion, plus multi-task learning, in terms of modeling the tangled semantic connections that show up inside internet memes.

3.2. Discussion

The achieved overall Macro F1-Score of 0.4278 for the complex, fine-grained multi-task classification of meme affect represents a significant contribution to the body of knowledge in multi-modal affective computing. This result must be interpreted in light of the prior known challenges established in the Introduction, particularly concerning sophisticated feature fusion and multi-task optimization.

3.2.1. Interpretation against Prior Knowledge

Previous works addressing the Memotion dataset [7, 12] often relied on computationally simpler fusion techniques such as Simple Early Fusion (Concatenation) or basic fixed Cross-Modal Feature Integration. These approaches inherently result in performance saturation because they fail to resolve the information conflict between modalities, where the image and text convey contradictory or only partially correlated signals. Our finding directly addresses the research gap regarding adaptive, dynamic fusion. The architecture's success validates the hypothesis that a dual-stage, dynamically-weighted mechanism is essential:

- **Cross-Modal Interaction:** The initial Bidirectional Cross-Attention step proved crucial for complex tasks like Sarcasm and Offensiveness. By allowing RoBERTa's linguistic features to contextualize ResNet's visual features (and vice versa), the model gained the necessary semantic depth to detect incongruity the key signal in satire which is often missed when features are simply concatenated.
- **Dynamic Weighting (GMF):** The introduction of the Gated Multi-modal Fusion (GMF) layer is the primary driver of the performance leap seen in Table II. The GMF layer automatically learned a sample-specific weighting mechanism, adapting to rely more heavily on the most salient feature. This dynamic weighting prevented noise from the less-relevant modality from degrading the fused representation, thereby boosting overall performance beyond established journal benchmarks like [14].

3.2.2. Contribution of Multi-Task Optimization

A core novelty of this research is the implementation of Uncertainty Weighting on the loss function, which successfully addressed the problem of negative transfer in MTL. As illustrated in the consistent convergence in Figure 4, the model efficiently balanced the highly skewed difficulty of the four tasks:

- The consistent training stability (Figure 4) and low standard deviation in the final score confirm that the model efficiently balanced the highly skewed difficulty of the four tasks. The adaptive σ^2 parameter allowed high-resource tasks (like Motivational classification) to support shared feature learning without dominating or degrading the performance of the more challenging, low-resource tasks (like Humor and Offensive classification), thereby maximizing knowledge transfer across the shared encoder layers.

- The results demonstrate that treating task difficulty as a learnable parameter significantly enhances robustness and generalization capacity compared to non-adaptive MTL models discussed in the literature.

3.2.3. Societal and Scientific Contribution

This work sort of contributes to multimodal meme understanding, by introducing MAGNUM, which is kind of an adaptive framework that mixes bidirectional cross-attention with Gated Multimodal Fusion so you can integrate textual info and visual cues in a more effective way. The architecture that is proposed, shows better results across several meme classification tasks, and it sort of underlines the usefulness of dynamic multimodal feature fusion, in practice. From a societal point of view, better performance especially in offensiveness detection, may help create more dependable content moderation systems. Also the fusion strategy itself gives a more flexible basis, that can be pushed forward into other multimodal classification setups, where there are tricky interactions between heterogeneous data sources.

4. CONCLUSION

In this paper, they proposed a multi-task deep learning system called MAGNUM (Multi-modal Attention and Gated Network for Understanding Memes) to address the main challenges in fine-grained meme categorization by emotion. The primary goal was to create a robust architecture that is able to handle the weaknesses common in previous approaches, such as those based on static feature fusion, and also fixed multi-task optimization choices. The results of the experimental showed that the proposed structure is indeed effective in demonstrating complex relationships between text and images required to understanding and comprehending memes. One key part of MAGNUM is the incorporation of a two-stage fusion strategy, created by Cross-Modal Attention and Gated Multimodal Fusion (GMF). The attention component is used to align characteristics across modalities, the GMF layer is used to re-weighted, and rebalance their effect in the fly depending on the characteristics of each individual sample. This is very important for models to be more consistent in handling noisy, incomplete or situational input, and is what adaptive fusion does. The improvements in performance suggest that there can be a more robust and faithful description of meme meaning when using adaptive multimodal fusion than when using the traditional static fusion techniques.

Furthermore, the unstable multi task learning was greatly facilitated by the uncertainty-based task weighting component: the model could sort of stabilize itself on which task is most important to learn in each classification task (Humor, Sarcasm, Offensive, and Motivational). That is, without us having to make any manual adjustments to that relative weight, it just changed in real-time, and it helped to reduce task interference without significantly altering the optimum during the entire learning process. Finally, the overall Macro F1-Score of the framework land was 0.4278 ± 0.0083 , indicating trustworthy behavior in all cross-validation folds. In a general sense, these results confirm that MAGNUM is a relatively good framework for multimodal meme analysis, since together these two work optimally: adaptive feature fusion and multi task learning. In addition to the advantages for the Memotion dataset, the proposed architecture will also provide a flexible platform for further exploration, including in multi-modal classification tasks where you require several visual input modalities, along with text.

4.1 Future Research Opportunities

The proposed MAGNUM framework provides a few, rather obvious, directions for further study. Specifically, there are a number of avenues that would be worthwhile to pursue further:

- Even if the GMF module combines multimodal information adaptively, future research could be guided by improving the visualization of the gating weights and attention distributions to gain insights into what is happening. The purpose is to understand that some predictions are made with the help of the visual and textual cues. We can make responsible content moderation applications more trustworthy and increase transparency if we can acquire it.
- The current approach uses mostly global visual representations for context modeling with a finer grainedness. To represent more complex interactions between picture areas and text, it is logical to include object-level signals together with region-based visual attention processes. On more challenging problems, such as offensive and humor classification, where signals may be somewhat localized, this could be quite helpful.
- It is to be noted that the version of this model that is being described here assumes that we have both text and images. It is assumed that we have both text and images to begin with; this is the version of the model being described. Therefore, modality aware training, or training procedures that take modality present into account, might be the subject of future studies. If either modality was not present

or was louder or shorter, the ideal is that the framework would still work well. As a result, it would be more appropriate to be used in a real-life social media environment, where perfection is rarely reached.

5. REFERENCES:

- [1] X. Luo, X. Chen, X. He, L. Qing, and X. Tan, "CMAFGAN: A Cross-Modal Attention Fusion based Generative Adversarial Network for attribute word-to-face synthesis," *Knowledge-Based Systems*, vol. 255, p. 109750, Nov. 2022, doi: 10.1016/j.knosys.2022.109750.
- [2] M. Xu, F. Liang, X. Su, and C. Fang, "CMJRT: Cross-Modal Joint Representation Transformer for Multimodal Sentiment Analysis," *IEEE Access*, vol. 10, pp. 131671–131679, 2022, doi: 10.1109/ACCESS.2022.3219200.
- [3] Y. Wang, Y. Xie, J. Zeng, H. Wang, L. Fan, and Y. Song, "Cross-modal fusion for multi-label image classification with attention mechanism," *Computers and Electrical Engineering*, vol. 101, p. 108002, July 2022, doi: 10.1016/j.compeleceng.2022.108002.
- [4] G. K, S. S. K, Y. H, V. M, and D. R, "DATA ANALYTICS SYSTEM FOR OFFENSIVE MEMES TEXT CLASSIFICATION IN SOCIAL NETWORK," *ShodhKosh J. Vis. Per. Arts*, vol. 5, no. 6, June 2024, doi: 10.29121/shodhkosh.v5.i6.2024.4162.
- [5] G. Xiang et al., "Driver multi-task emotion recognition network based on multi-modal facial video analysis," *Pattern Recognition*, vol. 161, p. 111241, May 2025, doi: 10.1016/j.patcog.2024.111241.
- [6] P. Adakane and A. Gaikwad, "Enhancing Multimodal Sentiment Prediction with Cross-Modal Attention and Adaptive Feature Weighting," *SSRG-IJECE*, vol. 12, no. 5, pp. 363–378, May 2025, doi: 10.14445/23488549/IJECE-V12I5P130.
- [7] N. Aslam et al., "MEDeep: A Deep Learning Based Model for Memotion Analysis," *MMEP*, vol. 9, no. 2, pp. 533–538, Apr. 2022, doi: 10.18280/mmep.090232.
- [8] N. Jindal, P. K. Kumaresan, R. Ponnusamy, S. Thavareesan, S. Rajiakodi, and B. R. Chakravarthi, "MISTRA: Misogyny Detection through Text–Image Fusion and Representation Analysis," *Natural Language Processing Journal*, vol. 7, p. 100073, June 2024, doi: 10.1016/j.nlp.2024.100073.
- [9] P. Meel and D. Kumar Vishwakarma, "Multi-modal fusion using Fine-tuned Self-attention and transfer learning for veracity analysis of web information," *Expert Systems with Applications*, vol. 229, p. 120537, Nov. 2023, doi: 10.1016/j.eswa.2023.120537.
- [10] T. Zhu, L. Li, J. Yang, S. Zhao, and X. Xiao, "Multimodal Emotion Classification With Multi-Level Semantic Reasoning Network," *IEEE Trans. Multimedia*, vol. 25, pp. 6868–6880, 2023, doi: 10.1109/TMM.2022.3214989.
- [11] Z. Quan, T. Sun, M. Su, and J. Wei, "Multimodal Sentiment Analysis Based on Cross-Modal Attention and Gated Cyclic Hierarchical Fusion Networks," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–12, Aug. 2022, doi: 10.1155/2022/4767437.
- [12] M. Velmala, S. Rajiakodi, K. Pannerselvam, and B. Sivagnanam, "Multimodal Sentiment Analysis of Online Memes: Integrating Text and Image Features for Enhanced Classification," *Procedia Computer Science*, vol. 258, pp. 355–364, 2025, doi: 10.1016/j.procs.2025.04.272.
- [13] Y. Duan, X. Zhao, and H. Guo, "Sentiment-aware cross-modal semantic interaction model for harmful meme detection," *Decision Support Systems*, vol. 196, p. 114509, Sept. 2025, doi: 10.1016/j.dss.2025.114509.
- [14] B. Wang, S. Huang, B. Liang, G. Tu, M. Yang, and R. Xu, "What do they 'meme'? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding," *Knowledge-Based Systems*, vol. 294, p. 111778, June 2024, doi: 10.1016/j.knosys.2024.111778.
- [15] Bologna Working Group on Qualifications Frameworks, A Framework for Qualifications of the European Higher Education Area. Last accessed Jun. 2015. [Online]. Available: http://www.bologna-bergen2005.no/Docs/00Main_doc/050218_QF_EHEA.pdf