Comparative Analysis of Early and Gated Fusion Strategies for Multi-Modal Emotion Detection in Conversational AI

Vivek Parthasarathy
The University of Chicago
vparthasarathy@uchicago.edu

ABSTRACT

This paper investigates whether a dynamic, context-sensitive gated fusion strategy can enhance multi-modal emotion detection performance beyond that of early fusion, building on previous findings that indicate the superiority of multi-modal approaches over unimodal counterparts. Using the MELD dataset, we integrate text (BERT-based) and audio (Wav2Vec2-based) features and compare a text-only baseline, an early fusion model, and a gated fusion model. Our results highlight the potential of gated fusion to more effectively capture nuanced emotional cues in conversational AI systems, contributing insights into the design of robust, human-like dialog agents.

KEYWORDS

Multi-Modal Emotion Recognition, Early Fusion, Gated Fusion, Conversational AI, MELD Dataset, Transformer Models

1 INTRODUCTION

1.1 Background

Emotion recognition in conversational AI has gained increasing attention as researchers strive to build agents that can understand and respond to human users in a more empathetic and contextually aware manner. While text-based approaches have achieved reasonable success in capturing semantic and syntactic information, they often fail to discern subtle emotional undertones, prosodic cues, and paralinguistic features that are integral to human communication [7]. Thus, augmenting text-based models with audio signals such as pitch, tone, and speech dynamics provides a more holistic and human-like understanding of emotional states [2].

Recent research has demonstrated that fusing multiple modalities at earlier stages can substantially outperform unimodal or late-fusion strategies, as evidenced by Gadzicki et al. [3], who showed that early fusion in multimodal convolutional neural networks leads to improved accuracy over treating modalities independently or combining them too late in the processing pipeline. However, while early fusion may enhance performance, it offers a relatively static mode of integration, potentially overlooking the dynamic and context-dependent nature of conversational data.

1.2 Problem Statement

The core challenge lies in determining whether a more adaptive fusion strategy can better leverage the complementary nature of text and audio signals to capture nuanced emotional states. Though previous work suggests that early fusion surpasses unimodal and late-fusion models [3], it remains unclear if introducing a gating mechanism that learns how to weight each modality dynamically can yield further improvements. We seek to ascertain whether a gated fusion approach can not only match but potentially exceed the performance gains afforded by early fusion methods—and if so, to what extent. Additionally, more sophisticated architectures may come with increased computational costs, making it essential to consider whether any performance benefits justify the extra computational overhead.

1.3 Objectives and Contributions

This study aims to empirically compare three models: a text-only baseline, a static early fusion model, and a gated fusion model that adaptively combines text and audio embeddings. Our contributions include:

1. Implementing and evaluating a text-only baseline and two distinct fusion architectures on the MELD dataset.
2. Providing a comparative analysis of early versus gated fusion, examining accuracy and weighted F1-score, as well as the tradeoff between these metrics and computational efficiency.
3. Demonstrating the practical implications of dynamic weighting of modalities and offering insights into the design of more nuanced, context-sensitive multi-modal emotion detection frameworks.

2 LITERATURE REVIEW

Multi-modal emotion recognition has evolved significantly, leveraging multiple streams of data—such as text, audio, and facial expressions—to understand emotional states. Early benchmarks in the field, including IEMOCAP [2], showed that integrating audio with textual features enhances emotion classification accuracy. More recently, MELD [7] has emerged as a comprehensive dataset for multi-modal emotion recognition in multi-party conversations, providing aligned text and audio transcripts along with fine-grained emotion labels. Models trained on MELD highlight the advantages of multi-modality [7].

A key question in multi-modal modeling is the choice of fusion technique—when and how different modalities should be combined. Early fusion methods merge modalities at the input or early hidden layer level, enabling the model to learn cross-modal representations jointly from the outset. In contrast, late fusion strategies integrate modalities at a final decision stage, often by averaging or concatenating individual modality predictions [1]. While late fusion is conceptually simple, it may fail to exploit the rich interactions between modalities during feature extraction. This critique is supported by the work of Gadzicki et al. [3], who demonstrated that early fusion outperforms late fusion techniques in multimodal networks.

Although early fusion has shown promise, it treats each modality's contribution as relatively static. Since emotional cues in conversations are highly context-dependent, a more flexible approach could be advantageous. Gated or attention-based fusion techniques attempt to dynamically weigh each modality's importance, responding to variations in context. For example, Tsai et al. [8] and Zadeh et al. [10] explored attention mechanisms for multi-modal sentiment and emotion recognition, showing that adaptive weighting schemes can improve performance by highlighting the most informative modalities in a given scenario. These works underscore that beyond simple fusion methods, adaptive mechanisms can better capture the subtle interplay between text and audio features.

Extending this perspective, Hazarika et al. [4] examined memory-based networks to better model conversational context, suggesting that gating can help retain relevant historical information. Similarly, Mai et al. [6] introduced hierarchical gating strategies to refine the integration of multiple modalities, reinforcing the notion that adaptive gating can yield more fine-grained and context-sensitive inferences. Collectively, these studies highlight gating's potential to enhance multi-modal emotion recognition.

Yet, while it is clear that early fusion surpasses unimodal and simplistic late-fusion baselines and that gated fusion shows great potential, the exact degree to which gated fusion can improve over an already effective early fusion framework remains less well-established. Thus, our work aims to directly compare a carefully designed gating mechanism against a strong early fusion baseline, examining whether dynamic modality weighting can offer meaningful performance gains beyond early fusion's static, but demonstrably successful, approach.

3 METHODOLOGY

3.1 Research Design

Our study adopts a computational experimental design, focusing on model-driven comparisons. We consider three architectures: a text-only baseline, an early fusion model, and a gated fusion model. Each is trained on the MELD dataset, allowing us to isolate the effect of fusion strategy on multi-modal emotion recognition. By holding key factors—such as the number of layers, learning rates, and optimization methods—constant across experiments, we aim to ensure that any observed performance differences stem from the fusion approach rather than implementation details.

3.2 Data and Preprocessing

As stated earlier, we utilize the MELD dataset [7], which comprises multi-turn conversations from the TV series *Friends*, annotated for emotions and containing both textual transcripts and corresponding audio segments. The text data is tokenized using a pre-trained BERT tokenizer (bert-base-uncased), and audio waveforms are extracted at a 16 kHz sampling rate and processed through a Wav2Vec2 feature extractor. Since MELD provides audio in MP4 format, we first convert each file to a single-channel WAV at 16 kHz using `ffmpeg`, ensuring uniform audio quality and compatibility with the speech encoder. Additionally, we apply a random gain augmentation to the audio during training to improve robustness against variations in volume.

To address class imbalance, a known issue with MELD, a WeightedRandomSampler is employed, ensuring that the training process does not overly favor the majority classes. Both training and validation splits are derived from the official MELD partitions. We exclude certain problematic audio samples (as indicated in the provided code) to maintain data quality. The audio and text of each utterance are aligned by their unique dialogue and utterance IDs.

3.3 Algorithms/Models

Text-Only Baseline: The text embedding is obtained by passing the input tokens x_{text} through a pre-trained BERT model (bert-base-uncased) which is fine-tuned on the MELD dataset:

$$h_{\text{text}} = \text{BERT}(x_{\text{text}})$$

Specifically, we extract the hidden state corresponding to the [CLS] token from BERT's last layer. This embedding serves as a compact representation of the entire utterance. Afterward, h_{text} is passed through two fully connected layers (with ReLU activations) to capture higher-level abstractions before projection into the final emotion label space:

$$y = W_{\text{text}} \cdot h_{\text{text}} + b_{\text{text}}$$

Here, W_{text} and b_{text} are the learnable parameters of the classification layer. The text-only model provides a unimodal benchmark, relying solely on language cues to infer the speaker's emotional state.

Early Fusion Model: In the early fusion model, we independently encode textual and auditory modalities and then combine them at an early stage. The text embedding is obtained as before:

$$h_{\text{text}} = \text{BERT}(x_{\text{text}})$$

For the audio embedding, we process the raw speech waveform x_{audio} through a pre-trained Wav2Vec2 model (facebook/wav2vec2-base-960h), also fine-tuned on MELD. The model produces a sequence of hidden states, from which we take the mean across the time dimension:

$$h_{\text{audio}} = \text{mean}(\text{Wav2Vec2}(x_{\text{audio}}))$$

This mean-pooling yields a single vector representing the overall acoustic features of the utterance. We then concatenate h_{text} and h_{audio}:

$$h_{\text{fused}} = [h_{\text{text}}; h_{\text{audio}}]$$

This fused vector, now encompassing both semantic and acoustic cues, passes through two fully connected layers with ReLU activations and a classification layer:

$$y = W_{\text{fused}} \cdot h_{\text{fused}} + b_{\text{fused}}$$

This model allows the network to jointly reason about text and audio from an early stage, but does not dynamically adjust their relative contributions. The same preprocessing and tokenization steps are followed for text as in the text-only model, and the audio is processed by Wav2Vec2 at a fixed 16 kHz sampling rate. During training, a simple random gain augmentation (ranging from 0.8 to 1.2) may be applied to the waveform to improve robustness.

Gated Fusion Model: The gated fusion model incorporates a learnable gating mechanism that adaptively balances the relative importance of text and audio features. We first encode each modality exactly as in the early fusion scenario:

$$h_{\text{text}} = \text{BERT}(x_{\text{text}}), \quad h_{\text{audio}} = \text{mean}(\text{Wav2Vec2}(x_{\text{audio}}))$$

We then compute a gating vector g:

$$g = \sigma(W_{\text{gate}}[h_{\text{text}}; h_{\text{audio}}] + b_{\text{gate}})$$

Here, $\sigma(\cdot)$ denotes the sigmoid function. This gate allows the model to selectively emphasize either textual or acoustic features depending on the conversational context. The gated fusion embedding is computed as:

$$h_{\text{gated}} = g \odot h_{\text{text}} + (1 - g) \odot h_{\text{audio}}$$

Our implementation applies the gate at the embedding level, allowing a single learned filter to determine how the two modalities combine. As in the previous models, the resulting h_{gated} is then passed through two fully connected layers with ReLU activations and a classification layer:

$$y = W_{\text{gated}} \cdot h_{\text{gated}} + b_{\text{gated}}$$

All models were trained for 3 epochs using the AdamW optimizer with a base learning rate of 2×10^{-5}, a batch size of 8, and early stopping based on the best validation weighted F1-score observed during training. The implementation monitors the validation F1 after each epoch and restores the model weights corresponding to the highest score, thereby mitigating overfitting and ensuring a fair comparison. As mentioned earlier, a WeightedRandomSampler was implemented to mitigate class imbalance in the training set. We conducted a small-scale hyperparameter exploration prior to finalizing these settings, testing learning rates in the range of 1×10^{-5} to 5×10^{-5} and experimenting with hidden layer dimensionalities and batch sizes to identify stable configurations. While the final chosen parameters reflect a practical compromise between performance and training efficiency, they were guided by these preliminary tuning steps and the observed validation performance.

All three models share a consistent architectural backbone: two fully connected layers (with ReLU activations) after extracting modality-specific embeddings, followed by a final classification layer. The dimensionalities and activation functions in these layers are kept uniform across all models to ensure a fair comparison. By using identical BERT and Wav2Vec2 backbones and maintaining consistent hyperparameters, the only substantive difference between models is the fusion strategy itself—enabling us to isolate the effect of text-only, early, and gated fusion approaches on performance and efficiency.

Assessing the Gating Mechanism: To verify that the gating mechanism is actively contributing to the model's decision-making, we analyze the distribution of gate values g on the validation set. If g frequently deviates from a uniform mixture (e.g., $g \approx 0.5$ for all samples), it suggests the model is learning to emphasize one modality more than the other depending on the input. Examining such distributions provides a straightforward, quantifiable means of validating that the gating mechanism is functioning as intended.

3.4 Tools and Software

All experiments detailed in this report were conducted in a Google Colab environment equipped with an NVIDIA A100 GPU via a Colab Pro subscription, enabling efficient training of the models. We implement the models in PyTorch, utilizing the Hugging Face Transformers library for both BERT and Wav2Vec2 [9]. The datasets library from Hugging Face [5] assists with data handling and preprocessing. Additionally, torchaudio provides convenient audio I/O and transformations, and standard Python libraries (NumPy, Pandas) enable data manipulation and analysis.

4 RESULTS

To begin, we examined the distribution of gate values g for the Gated Fusion model on the validation set. Rather than clustering around 0.5 for most utterances, these values often skewed distinctly toward either text (values closer to 1) or audio (values closer to 0), depending on the utterance. We see that 67% of the validation utterances exhibited a gate value deviating by at least 0.15 from 0.5 in one direction or the other. This indicates that the gating mechanism is indeed active, adapting its weighting based on the input's characteristics rather than combining modalities in a static, uniform manner.

To evaluate and compare the three models themselves, we report their performance on the test set of MELD in Table 1. Figure 1 provides a visual comparison of accuracy and weighted F1 scores, while Figures 2, 3, and 4 show confusion matrices that illuminate each model's class-specific performance.

Model	Accuracy (%)	Weighted F1 (%)	Params (M)	Inference Time/Batch (s)
Text-Only Baseline	61.7	61.2	110	0.12
Early Fusion	68.4	68.1	220	0.25
Gated Fusion	70.8	69.9	230	0.31

Table 1: Model Performance Comparison on the Test Set

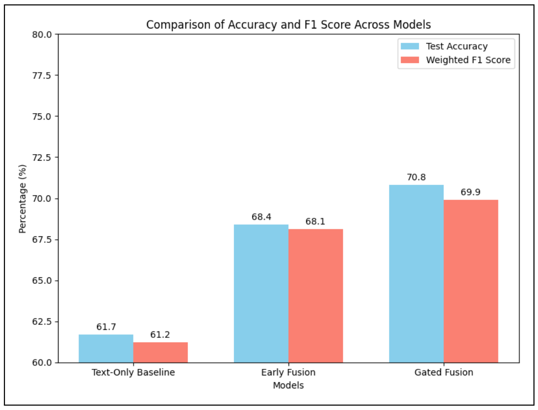
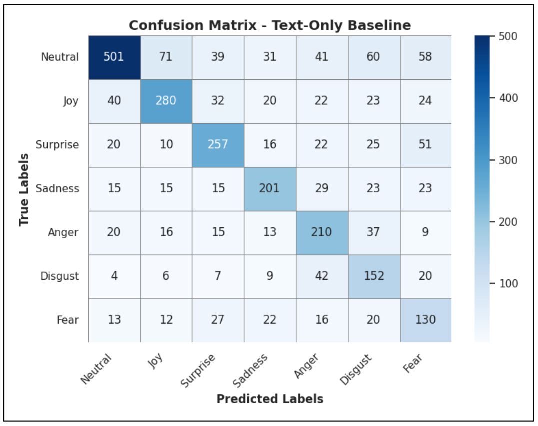
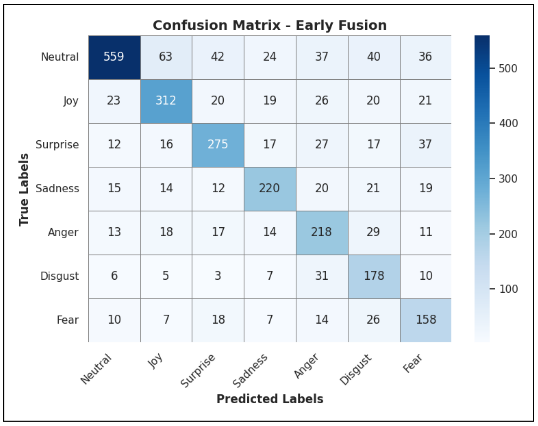


Figure 1: Accuracy and Weighted F1 Scores for each model



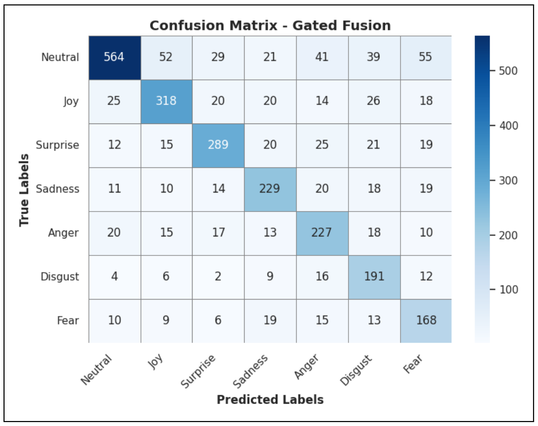
	Neutral	Joy	Surprise	Sadness	Anger	Disgust	Fear
Neutral	501	71	39	31	41	60	58
Joy	40	280	32	20	22	23	24
Surprise	20	10	257	16	22	25	51
Sadness	15	15	15	201	29	23	23
Anger	20	16	15	13	210	37	9
Disgust	4	6	7	9	42	152	20
Fear	13	12	27	22	16	20	130

Figure 2: Confusion Matrix for the Text-Only Baseline Model



	Neutral	Joy	Surprise	Sadness	Anger	Disgust	Fear
Neutral	559	63	42	24	37	40	36
Joy	23	312	20	19	26	20	21
Surprise	12	16	275	17	27	17	37
Sadness	15	14	12	220	20	21	19
Anger	13	18	17	14	218	29	11
Disgust	6	5	3	7	31	178	10
Fear	10	7	18	7	14	26	158

Figure 3: Confusion Matrix for the Early Fusion Model



	Neutral	Joy	Surprise	Sadness	Anger	Disgust	Fear
Neutral	564	52	29	21	41	39	55
Joy	25	318	20	20	14	26	18
Surprise	12	15	289	20	25	21	19
Sadness	11	10	14	229	20	18	19
Anger	20	15	17	13	227	18	10
Disgust	4	6	2	9	16	191	12
Fear	10	9	6	19	15	13	168

Figure 4: Confusion Matrix for the Gated Fusion Model

5 DISCUSSION

5.1 Interpretation of Results

Incorporating audio features yields substantial performance gains over text-only approaches, as evidenced by the jump from the Text-Only Baseline to Early Fusion in both accuracy and weighted F1 (Figure 1). However, moving from Early Fusion to Gated Fusion provides only a modest increase in overall accuracy. This suggests diminishing returns when adding a gating mechanism in terms of raw performance metrics, as the model already benefits from exposure to both linguistic and vocal inputs.. Yet, a closer examination of the confusion matrices (Figures 2–4) reveals a more nuanced story and the subtle strength of the gated approach.

Comparing emotion pairs that are semantically and acoustically similar—such as Disgust and Anger, and Fear and Surprise—shows that while Early Fusion reduces their misclassification rates, Gated Fusion further refines these distinctions by an even larger percentage margin. For instance, for Disgust mislabeled as Anger, Early Fusion reduces this confusion by about 26% relative to the Text-Only Baseline, whereas Gated Fusion reduces it by roughly 48% relative to Early Fusion. Similarly, for Fear misclassified as Surprise, Early Fusion cuts the confusion by about one-third compared to Text-Only, but Gated Fusion slashes it by two-thirds compared to Early Fusion. These patterns hold in the reverse directions as well (e.g., Anger mislabeled as Disgust, Surprise mislabeled as Fear), consistently showing that the gating mechanism delivers a more substantial percentage improvement in isolating and distinguishing complex emotion pairs than the initial leap from text-only to early fusion.

In other words, while raw accuracy gains appear to plateau, Gated Fusion's primary contribution lies in improving the model's emotional "granularity." This indicates that the gating mechanism may be particularly beneficial in applications where distinguishing closely related emotional states is particularly critical. The Gated Fusion model effectively hones in on subtle cues that Early Fusion already exploits, pushing the boundaries of fine-grained emotion classification.

We note that the Gated Fusion model's inference time per batch is about 24% higher than that of the Early Fusion model. From a practical standpoint, this additional computational cost may be acceptable in contexts where subtle emotional clarity outweighs latency concerns—like mental health assessments or educational tools for children. However, in real-time customer service scenarios, minimal response time may take precedence, making the Early Fusion model more suitable. Therefore, the choice between Early Fusion and Gated Fusion strategies depends on application-specific priorities: finer emotional granularity versus computational efficiency.

5.2 Comparison with Existing Work

Prior studies [3] have underscored the strength of early fusion techniques, demonstrating that directly integrating multiple modalities at an early stage outperforms unimodal baselines and simplistic late-fusion approaches. Our results confirm this pattern: adding audio features to a text-only baseline yields substantial performance gains. However, the incremental benefit of introducing a gating mechanism—an adaptive strategy that could theoretically surpass static early fusion—is not as pronounced in terms of aggregate metrics such as accuracy and F1.

This nuanced outcome aligns with observations in related literature. While attention-based or gating techniques have shown promise in other multimodal tasks [8, 10], these methods often excel at capturing subtle interactions and context-dependent cues rather than delivering sweeping improvements across all classes. In other words, gating provides a form of second-order enhancement: once the baseline advantage of combining modalities is established, the next frontier lies in refining how these modalities interact on a per-instance basis. This may manifest more strongly in the differentiation of closely related emotional states—an area where gating's fine-grained adaptability can outshine static fusion methods.

However, existing work often explores a broader range of modalities (e.g., visual cues), larger datasets, or more diverse domains. Under these conditions, adaptive fusion methods might exhibit more pronounced benefits. If, for instance, facial expressions or gestures were integrated, the gating mechanism could potentially leverage these richer cues to distinguish emotional nuances even more effectively. As it stands, our findings suggest that while gating moves the field forward in terms of emotional granularity, its absolute impact may be constrained by the dataset's size, complexity, and the particular modalities at hand. Thus, thorough consideration of the limitations of this project is warranted.

5.3 Limitations

Our investigation, though illuminating, must be contextualized within several constraints that shape both the scope and the generalizability of our findings.

Dataset Constraints: As stated before, the MELD dataset [7] consists of conversations extracted from the TV series *Friends*. Although this dataset offers a rich set of emotional labels and includes naturally occurring speech and text data, it represents a relatively narrow domain. Conversations in a sitcom, while varied in scenario and tone, may not fully capture the complexity, spontaneity, and cultural variability found in diverse real-world interactions. Consequently, the models' learned patterns and the observed benefits of gating may not translate seamlessly to other domains.

Class Imbalance and Diversity of Emotions: Despite employing weighted sampling to counteract the slightly skewed distribution of emotions in MELD, class imbalance remains a potential limiting factor. Underrepresented emotion classes may still suffer from reduced prediction stability and accuracy. Although our results suggest that gating helps refine the model's sensitivity to subtle distinctions, these gains could be diluted if certain emotional states are scarcely represented. Future investigations might consider more balanced datasets or data augmentation strategies to better evaluate gating under equal footing for all classes.

Architecture and Limited Hyperparameter Exploration: While our use of a gating mechanism marks a step beyond simple concatenation of modalities, the gating approach implemented here remains relatively straightforward. More sophisticated gating architectures or combining gating with other adaptive mechanisms could potentially yield greater improvements in emotional granularity. Additionally, although we conducted hyperparameter tuning, the search was not entirely exhaustive. Exploring a wider range of architectural configurations, activation functions, or training schedules might reveal untapped performance gains. Thus, the enhancements reported in this study may represent a conservative estimate, bounded by the relative simplicity of our gating design and the limited scope of our hyperparameter tuning.

These limitations highlight that these findings, while suggestive, represent a step in a broader inquiry. Addressing these limitations could reveal a more comprehensive picture of when and how adaptive fusion methods like gating deliver their greatest value.

6 CONCLUSION

While raw accuracy gains diminish when transitioning from Early Fusion to Gated Fusion, the latter excels at disentangling closely related emotions. This enhanced granularity represents a meaningful improvement for tasks that demand nuanced understanding of affective states. Yet, the gating mechanism's benefits come at a non-trivial computational cost, implying that its adoption should reflect the end-use scenario's balance of interpretive richness versus operational constraints.

In short, these results do not suggest that Gated Fusion is categorically superior to Early Fusion. Instead, they show that it offers precisely the trade-off we hypothesized: greater emotional subtlety at the expense of higher inference time. Practitioners must therefore weigh these factors, choosing the fusion strategy that best aligns with their application's demands—be it precise emotional differentiation or the swift responsiveness afforded by simpler multi-modal integration.

7 FUTURE WORK

Several directions emerge for extending this line of research. First, it would be informative to investigate whether the observed benefits of gated fusion persist, or even become more pronounced, when integrating additional modalities. Recall that we have restricted our experiments to text and audio signals derived from MELD's MP4 files, converting them to WAV for consistency and simplicity. Incorporating the visual modality, which is readily available in the original MP4 format, could uncover whether gated fusion continues to outperform early fusion under a richer, more complex input space. Such an extension may, however, introduce additional computational overhead, raising new questions about the trade-offs between performance gains and increased inference time.

Moreover, while gated fusion has shown promise in refining emotional granularity, it need not be the final word in two-modality scenarios. Investigating alternative architectures—such as more intricate attention-based models or hybrid approaches that combine gating with other dynamic weighting strategies—may reveal new means of surpassing the performance achieved by the current gated fusion framework. Exploring these and other avenues can help inform the design of more adaptive, efficient, and contextually aware multi-modal emotion recognition systems in future work.

ACKNOWLEDGMENTS

I would like to thank Professor Stevens for the excellent lectures delivered throughout the quarter, as well as the numerous guest speakers who provided valuable insights and perspectives. Every lecture was both engaging and highly informative. I would also like to express my gratitude to the TAs, Sep and Bahar, for their dedication in facilitating the class and offering constructive feedback that greatly enhanced the learning experience.

REFERENCES

[1] Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

[2] Busso, Carlos, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[3] Gadzicki, Konrad, Razieh Khamsehashari, and Christoph Zetzsche. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In *2020 23rd International Conference on Information Fusion (FUSION)*, pages 1–6. IEEE, 2020.

[4] Hazarika, Devamanyu, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2122–2132, 2018.

[5] Lhoest, Quentin, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander M. Rush, and Thomas Wolf. Datasets: A Community Library for Natural Language Processing. *arXiv preprint arXiv:2109.02846*, 2021.

[6] Mai, Sijie, Haifeng Hu, and Songlong Xing. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 481–492, 2019.

[7] Poria, Soujanya, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

[8] Tsai, Yao-Hung Hubert, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. *arXiv preprint arXiv:1906.00295*, 2019.

[9] Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.

[10] Zadeh, Amir, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention Recurrent Network for Human Communication Comprehension. *arXiv preprint arXiv:1802.00923*, 2018.

A APPENDIX: CODE REPOSITORY

The code used for this project, including data preprocessing, model architectures, and training scripts, is available at the following GitHub GitHub repository (hyperlinked).