Multimodal Sentiment Analysis

Akshay Paralikar[*], **Rugved Mhatre**[*], **Sakshee Sawant**[*]

Tandon School of Engineering, New York University, New York, USA
{akshay.paralikar, rugved.mhatre, ss18198}@nyu.edu
GitHub: rugvedmhatre/Multimodal-Sentiment-Analysis

Abstract

Multimodal Sentiment Analysis (MSA) is a burgeoning research area that aims to understand human sentiments by leveraging information from multiple modalities, such as text, audio, and video. In recent years, the proliferation of social media platforms and multimedia content has necessitated the development of advanced computational models capable of analyzing sentiment expressed across various modalities. This paper explores different methodologies for MSA, focusing on fusion techniques that integrate information from diverse modalities to enhance sentiment analysis performance, and the various architectures that employ them. We investigate Early Fusion, Late Fusion, Tensor Fusion, and their variants, along with other approaches like Multimodal Factorization Model (MFM), Multimodal Cyclic Translation Network (MCTN), and Multimodal Transformer (MulT). Each approach offers unique advantages and challenges, providing diverse tools to tackle the complexities of sentiment analysis in multimodal data. Through a comprehensive review and analysis of these methodologies, this paper aims to shed light on the current state-of-the-art in MSA.

Introduction

Sentiment analysis, also known as opinion mining, is a subfield of natural language processing that aims to identify, extract, and quantify subjective information from text data. Over the years, sentiment analysis has evolved from analyzing textual data alone to encompassing multiple modalities, including audio, video, and images. This expansion is driven by the increasing availability of multimedia content on social media platforms, online forums, and digital communication channels. MSA leverages information from various modalities to gain a deeper understanding of human emotions, attitudes, and opinions expressed in multimedia content.

The integration of multiple modalities poses both challenges and opportunities for sentiment analysis. While text-based sentiment analysis techniques have matured significantly, analyzing sentiments expressed in audio, video, and other modalities requires specialized models and feature fusion techniques. Fusion techniques combine information from different modalities to enhance the overall sentiment analysis performance. In this paper, we provide a comprehensive review of fusion techniques for MSA, focusing on their methodologies, advantages, and limitations. By examining the latest research developments in multimodal architectures, we aim to provide insights into the current state-of-the-art and identify future research directions in this exciting field.

This paper makes the following contributions:

- Explanation of the taxonomy being used in Multimodal learning research
- Exploration of various MSA architectures - Early Fusion, Late Fusion, Tensor Fusion, Low Rank Tensor Fusion, MFM, MCTN, and MulT.
- Evaluation on CMU-MOSI and CMU-MOSEI datasets.

Literature Survey

Multimodal sentiment analysis has garnered significant attention in recent years. Researchers have explored different approaches to fuse information from multiple modalities to improve sentiment analysis performance. The work by Lai et al. (2023) lays out the landscape of MSA. It defines the field, examines recent datasets, and explores various models. Our project draws heavily from this survey, using it to choose datasets and model architectures.

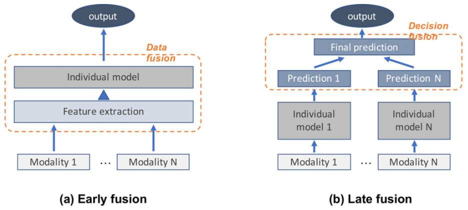

Figure 1: Early Fusion Vs. Late Fusion

In MSA, feature fusion techniques like Early Fusion and Late Fusion are commonly employed (Baltrušaitis, Ahuja, and Morency 2017). Early fusion integrates features right after extraction, often by concatenating their representations. Conversely, Late Fusion integrates features after each modality has made a decision, such as classification or re-

[*] These authors contributed equally.

gression (Figure 1). In our paper, we delve into this approach by employing GRU and Transformer models.

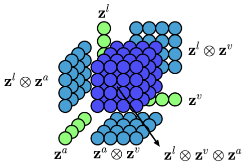
$\mathbf{z}^l$
$\mathbf{z}^l \otimes \mathbf{z}^v$
$\mathbf{z}^l \otimes \mathbf{z}^a$
$\mathbf{z}^v$
$\mathbf{z}^a$
$\mathbf{z}^a \otimes \mathbf{z}^v$
$\mathbf{z}^l \otimes \mathbf{z}^v \otimes \mathbf{z}^a$

Figure 2: Tensor Fusion

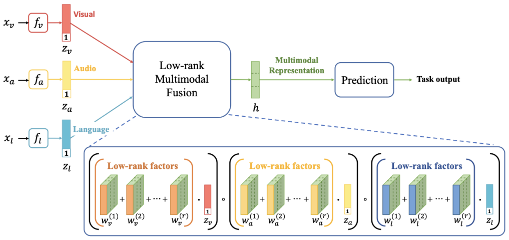

Figure 3: Low Rank Tensor Fusion Model

Tensor fusion techniques (Figure 2) have been explored to capture interactions between modalities effectively. We implement Zadeh et al. (2017) Tensor Fusion model to learn the intra-modality and inter-modality dynamics. We also explore an efficient improvement, by Liu et al. (2018), the Low Rank Tensor Fusion model (Figure 3).

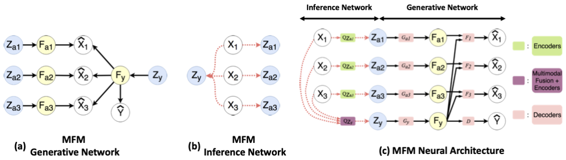

Figure 4: MFM with three modalities

Tsai et al. (2019b) introduce a Multimodal Factorization Model (MFM) model that dissects representations into two discernible sets: multimodal discriminative factors and modality-specific generative factors. The multimodal discriminative factors, common across all modalities, encapsulate shared features crucial for tasks such as sentiment prediction. In contrast, the modality-specific generative factors are distinct to each modality, encoding information vital for data generation. Our study integrates MFM architecture (Figure 4) to leverage modality-specific and shared representations for improved sentiment analysis performance.

Pham et al. (2020) propose Multimodal Cyclic Translation Network (MCTN) for learning robust joint representations through modality translation. MCTNs leverage modality translation to build joint representations solely from the source modality, making the model more resilient to missing or distorted information in other modalities. This approach represents a significant improvement over traditional fusion techniques. We investigate MCTNs (Figure 5) to understand their effectiveness compared to standard fusion methods and their potential for enhancing sentiment analysis tasks across different domains.

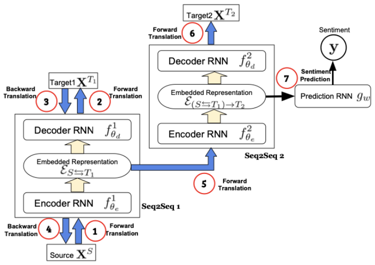
Target2 $\mathbf{X}^{T_2}$
Forward Translation
6
Sentiment
$\mathbf{y}$
Target1 $\mathbf{X}^{T_1}$
Backward Translation
3
2
Forward Translation
Decoder RNN $f_{\theta_d}^2$
Embedded Representation $\mathcal{E}_{(S \leftrightarrows T_1) \to T_2}$
Encoder RNN $f_{\theta_e}^2$
7
Sentiment Prediction
Prediction RNN g_w
Decoder RNN $f_{\theta_d}^1$
Embedded Representation $\mathcal{E}_{S \leftrightarrows T_1}$
Encoder RNN $f_{\theta_e}^1$
Seq2Seq 2
5
Forward Translation
Seq2Seq 1
Backward Translation
4
1
Forward Translation
Source $\mathbf{X}^S$

Figure 5: Hierarchical MCTN for three modalities

Finally, we implement the MulT model proposed by Tsai et al. (2019a), which introduces a directional pairwise crossmodal attention mechanism. This mechanism dynamically focuses on interactions among multimodal sequences across various time intervals and adeptly adjusts streams from one modality to another. Notably, the MulT model (Figure 6) surpasses existing state-of-the-art methods in MSA. Moreover, the crossmodal attention mechanism within MulT demonstrates effectiveness in capturing correlated cross-modal signals, underscoring its utility in capturing nuanced interactions between different modalities for sentiment analysis tasks.

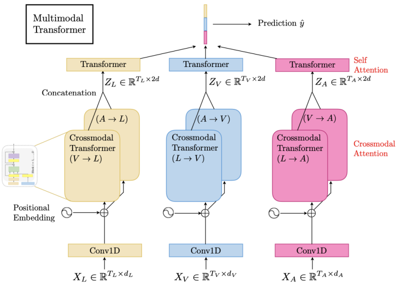
Multimodal Transformer
Prediction $\hat{y}$
Transformer
Transformer
Transformer
Self Attention
$Z_L \in \mathbb{R}^{T_L \times 2d}$
$Z_V \in \mathbb{R}^{T_V \times 2d}$
$Z_A \in \mathbb{R}^{T_A \times 2d}$
Concatenation
$(A \to L)$
$(A \to V)$
$(V \to A)$
Crossmodal Transformer $(V \to L)$
Crossmodal Transformer $(L \to V)$
Crossmodal Transformer $(L \to A)$
Crossmodal Attention
Positional Embedding
Conv1D
Conv1D
Conv1D
$X_L \in \mathbb{R}^{T_L \times d_L}$
$X_V \in \mathbb{R}^{T_V \times d_V}$
$X_A \in \mathbb{R}^{T_A \times d_A}$

Figure 6: MulT architecture for three modalities

We leverage the MultiBench tool developed by Liang et al. (2021), which offers a streamlined data loading pipeline and a comprehensive library of frequently used models. This resource facilitates our exploration of intricate model architectures, as previously discussed, by simplifying the experimentation process. By harnessing MultiBench, we delve into more complex model implementations for MSA, thus enriching our understanding of effective approaches in this domain.

Dataset

We utilize two datasets: **CMU-MOSI** (Zadeh et al. 2016) and **CMU-MOSEI** (Bagher Zadeh et al. 2018), which serve as foundational resources for MSA. The CMU-MOSI dataset comprises 2,199 opinion video clips meticulously annotated with various attributes, including subjectivity, sentiment intensity, and detailed visual and audio features. Notably, sentiment intensity spans a range from -3 to +3, enabling nuanced sentiment analysis. On the other hand, CMU-MOSEI offers a comprehensive platform for sentence-level sentiment analysis and emotion recognition in online videos. Spanning over 65 hours of annotated video material from diverse speakers and topics, it encompasses annotations for sentiment, nine discrete emotions, and continuous emotional dimensions such as valence, arousal, and dominance. The dataset's diverse tasks render it invaluable for evaluating and testing multimodal models in the realm of affective computing.

Data Preprocessing

The datasets are sourced from CMU's servers and preprocessed. Data preprocessing entails several steps to harmonize modalities and extract features. Initially, Glove word embeddings transform word sequences from video segment transcripts into 300-dimensional vectors, synchronized using the P2FA forced aligner to ensure alignment across text, audio, and video. Visual features, including facial action units and landmarks, are extracted via the Facet library and OpenFace tool, forming sequences that depict facial expressions over time. These visual cues are complemented by audio features obtained from COVAREP, capturing tone variations with 12 Mel-frequency cepstral coefficients and other relevant parameters. Ultimately, the resulting sequences encapsulate nuanced aspects of language, visual cues, and acoustic characteristics across the dataset's videos, providing a rich foundation for multimodal sentiment analysis.

Methodology

Initialization

The initialization procedure involves several key steps: initializing three encoder models tailored to process input data from specific modalities like text, audio, and video; instantiating a fusion module to merge the encoder outputs, creating a unified representation of multimodal input data; initializing a multimodal head to process the fused representation for final prediction or classification; and configuring the optimizer with appropriate hyperparameters like learning rate and weight decay to govern the optimization process. Adhering to this standardized initialization ensures consistency and facilitates the integration of various fusion strategies and architectures into the multimodal learning framework.

Early Fusion

We begin by implementing a classical multimodal architecture with the Early Fusion technique. Our model incorporates Gated Recurrent Unit (GRU) encoders and a MLP as the multimodal head. Utilizing the Adam optimizer with a learning rate of 0.001 and a weight decay of 0.01 facilitates model optimization. The choice of the L1 Loss function, suitable for regression tasks, guides the model to minimize the absolute difference between predicted and ground truth values. Despite its simplicity, this approach yields a test accuracy of 65.74% for CMU-MOSI and 49.03% for CMU-MOSEI, suggesting that the simplest approach may not always yield the best results.

Early Fusion with Transformer We enhanced the early fusion method by integrating Transformer models to handle sequential data more efficiently. This upgrade entails replacing the GRU layer with a Transformer layer, exploiting its attention mechanism to capture long-range dependencies within the fused representation. During training, we adjusted the learning rate to 0.0001 and implemented early stopping to mitigate overfitting risks. Nevertheless, the fundamental methodology, encompassing data preprocessing, fusion strategy, and evaluation metrics, remains aligned with the traditional early fusion approach. As a result of these enhancements, we observed improved test accuracy, achieving 76.96% for CMU-MOSI and 69.09% for CMU-MOSEI.

Late Fusion

Next, we contrast the aforementioned models with a Late Fusion approach. We employ a GRU encoder with an MLP head. We utilize the Adam optimizer with a learning rate of 0.001, a weight decay of 0.01, and the L1 Loss function. Surpassing the outcomes of simple early fusion, this method yields a test accuracy of 70.26% on CMU-MOSI.

Late Fusion with Transformer Similar to our previous experiments, we substitute the GRU layer with a Transformer layer. We maintain the same optimizers and loss function, albeit reducing the learning rate to 0.0001, while also implementing early stopping to address overfitting concerns. Consequently, this adjustment yields an improved test accuracy of 74.34% for our CMU-MOSI dataset.

Tensor Fusion

In this approach, a Tensor Fusion technique was employed to merge multimodal data streams from GRU encoders. The fused representation obtained from the Tensor Fusion module was then passed through a MLP multimodal head. Optimization was performed using the Adam optimizer with a learning rate of 0.001, a weight decay of 0.01, and the L1 Loss function. Surpassing the performance of the previous two approaches without incorporating a Transformer, this model achieved a test accuracy of 72.74% on CMU-MOSI and 67.11% on CMU-MOSEI.

Low-Rank Tensor Fusion In this method, we incorporated a Low-Rank Tensor Fusion (LRF) module to merge multimodal data streams. The LRF strategy entails projecting input features from each modality into a low-rank tensor space, followed by tensor fusion to generate a joint representation. This approach is notably more efficient compared to the previously mentioned Tensor Fusion method. We employed the Adam Optimizer with a learning rate of 0.001 and weight decay of 0.01, utilizing the L1 Loss function. However, despite its efficiency, this approach yielded inferior results compared to Tensor Fusion, achieving a test accuracy of 68.07% on CMU-MOSI.

Multimodal Factorized Model (MFM)

In this approach, we devised a model architecture depicted in Figure 4. It comprises modality-specific encoders and decoders, alongside a fusion module and a multimodal head.

The decoders are tasked with reconstructing the original input data from the latent representations produced by the encoders. Throughout the training phase, the MFM model underwent training using a multi-task objective function. This function amalgamated the reconstruction losses from the decoders and the classification loss from the multimodal head. Mean squared error loss computed the reconstruction losses, while cross-entropy loss quantified the classification loss. Training employed an RMSProp optimizer with a learning rate of 0.001. However, this model demonstrated overfitting, achieving 71.02% accuracy on CMU-MOSI training data and only 66.47% on the testing data.

Multimodal Cyclic Translation Network (MCTN)

The MCTN architecture comprises two key components: encoders and decoders (Figure 5). Similar to MFM, three modality-specific decoders were initialized, corresponding to each encoder, aimed at reconstructing the original input data from the latent representations generated by the encoders. During training, the MCTN model underwent a cyclic translation process, wherein encoders and decoders iteratively learned. Encoders generated latent representations, while decoders reconstructed the original input data from these representations. This cyclic training aimed to capture shared information across modalities and learn joint representations. The model employed various loss functions, including mean squared error loss for reconstruction tasks and L1 loss for regression tasks. Training utilized the Adam optimizer with a learning rate of 0.0001 and weight decay of 0.01. This approach yielded an accuracy of 73.76% on CMU-MOSI test data.

Multimodal Transformer Model (MulT)

Finally, we deploy the Multimodal Transformer for Unaligned Multimodal Language Sequences. MulT is equipped with a directional pairwise crossmodal attention mechanism, enabling it to capture interactions among different modalities across diverse time steps. This attention mechanism facilitates the adaptive fusion of information from multiple modalities while effectively handling misalignments between them. We adopt the architectural design illustrated in Figure 6. For optimization, we employ the Adam optimizer with a learning rate set to 0.001, a weight decay of 0.01, and utilize the L1 loss function. Our experimental results demonstrate the effectiveness of MulT, yielding promising metrics. We achieve an accuracy of 75.07% on the CMU-MOSI test dataset and 71.91% on the CMU-MOSEI test dataset.

Conclusion

In conclusion, our research delves into the realm of multimodal architectures and fusion techniques to enhance sentiment analysis. Through extensive experimentation, we observe contrasting performance trends: while simplistic models exhibit ease of training but limited accuracy, complex models offer improved accuracy albeit with slower computational speed. However, amidst this spectrum of approaches, Transformers emerge as the premier choice for sentiment analysis, owing to their powerful self-attention mechanisms. This finding underscores the significance of exploring transformer-based methodologies further. Future endeavors could involve investigating alternative techniques such as gradient blending and beyond, to continually refine and advance the field of multimodal sentiment analysis.

Model	Accuracy
Early Fusion (GRU)	65.74%
Early Fusion (Transformer)	76.96%
Late Fusion (GRU)	70.26%
Late Fusion (Transformer)	74.34%
Tensor Fusion	72.74%
Low Rank Tensor Fusion	68.07%
MFM	66.47%
MCTN	73.76%
MulT	75.07%

Table 1: Accuracies on CMU-MOSI

References

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Association for Computational Linguistics*.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2017. Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406.

Lai, S.; Hu, X.; Xu, H.; Ren, Z.; and Liu, Z. 2023. Multimodal Sentiment Analysis: A Survey. arXiv:2305.07611.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. https://github.com/pliang279/MultiBench/. arXiv:2107.07502.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. arXiv:1806.00064.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Poczos, B. 2020. Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. arXiv:1812.07809.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019a. Multimodal Transformer for Unaligned Multimodal Language Sequences. arXiv:1906.00295.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019b. Learning Factorized Multimodal Representations. arXiv:1806.06176.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv:1707.07250.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. arXiv:1606.06259.