National College of Ireland

Evaluation of Multimodal Transformer Data Fusion Techniques

MSc Research Project
MSCAI

David Oluwatimilehin Bamikole
Student ID: X22179640

School of Computing
National College of Ireland

Supervisor: Dr. Devanshu Anand

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	David Oluwatimilehin Bamikole
Student ID:	X22179640
Programme:	MSCAI
Year:	2024
Module:	MSc Research Project
Supervisor:	Dr. Devanshu Anand
Submission Due Date:	12/08/2024
Project Title:	Evaluation of Multimodal Transformer Data Fusion Techniques
Word Count:	4871
Page Count:	20

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	David Oluwatimilehin Bamikole
Date:	16th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Evaluation of Multimodal Transformer Data Fusion Techniques

David Oluwatimilehin Bamikole
X22179640

Abstract

Good context provides better insight into understanding a message, and this can be obtained by extracting information from different mediums through which the message is passed. In cases where the medium used does not provide a complete insight to best understand the message, assumptions are generated based on the extracted context or the message is left un-understood, both of which do not lead to good comprehension of the message. This also applies to the use of a single modal data such as text, audio or video for machine learning tasks. However, using multiple modalities requires the fusion of data from different modalities. There are existing data fusion strategies such as feature-level, decision-level and hybrid fusion approaches, all of which produce different levels of effectiveness along with several corresponding attributes. This resulted in this research work where data fusion techniques for multimodal transformers were evaluated. The CMU-MOSI dataset which has audio, textual and visual modalities was used to implement early concatenation fusion, cross-modal attention fusion and hierarchical modal attention fusion. The best hyperparameter was obtained for each strategy. Using the mean absolute error (MAE), Pearson coefficient correlation, parameter size and training time to evaluate the performance of the models, the hierarchical model performs best with 0.0111 MAE and 0.5509 coefficient score but also the largest and slowest model. The cross-modal transformer has the smallest parameter size and the early concatenation fusion has the fastest speed.

Keywords: multimodal transformer, data fusion techniques, early concatenation fusion, cross-modal attention fusion, hierarchical modal attention fusion.

1 Introduction

Humans use multiple communication channels, which explains the complexity of human communication. Humans communicate through speech, expressions, signs and writings, these channels are known as modalities. A message can be passed through any of the channels, and the context of the message passed is enriched through other channels. Human speech is always accompanied by gestures, and the use of gestures can significantly alter the meaning of the speech. In addition, the McGurk effect (Flores-Coronado et al., 2022) has proven that using multiple channels for communication could influence the message received by the recipient. In the experiment, different audio-visual stimuli were paired and it resulted in a different syllable being perceived, a video of a person uttering /ga/ underlay with a speech of /ba/ was perceived as /da/. However, it was noticed that recipients who were not viewing the video but heard only the sound perceived

the right utterance. This experiment shows the impact of using multiple modalities for communication.

However, the problem of utilizing multimodal AI systems over unimodal AI systems persists. It could be observed that vast existing AI systems are unimodal, using only one modality for communication with the common modality being text. Although recent research has affirmed the positive impact of multimodal systems over unimodal systems (Li et al., 2024), most research focuses on two modalities such as textual-visual pair modalities and textual-audio pair modalities, neglecting the third common modality for basic human communication which is usual audio modality or visual modality (Taheri et al., 2023).

In addition, the challenge of focusing on a single modality could be seen in the state-of-the-art transformer model. The design of transformer model (Vaswani et al., 2017) which is text-based led to several models such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2019) which all earn state-of-the-art, although all textual based. Different variants of the text-based model emerge for other modalities such as Vision transformer (Dosovitskiy et al., 2020) and Visual-BERT (Su et al., 2019) for visual modality and Wav2Vec (Baevski et al., 2020) for audio modality.

However, using unimodal models for multimodal tasks led to the problem of harmonization strategies for multiple modalities. Existing solutions to this challenge use feature-level, decision-level and hybrid fusion approaches. These approaches involved using multiple unimodal models to process the different modalities and combining the extracted feature or/and decision of each modality model to obtain the final result (Karani and Desai, 2022). However the transformer model had a unique attribute in its architecture, the multi-head attention layer which helps it focus on different sections of the input data. This has been explored for developing multimodal transformers over the existing unimodal transformer, and this has also resulted in new data harmonization strategies (Xu et al., 2023).

The motivation for this research work was derived from the neglect of using the three basic modalities (audio, textual and visual) in human communication for multimodal systems and fusion strategy for multimodal transformers. The various data harmonization strategies for multimodal transformers come with different impact which requires attribute trade-offs depending on the purpose it is being required.

1.1 Research Question

What is the effectiveness and impact of various multimodal transformer data fusion strategies and the trade-off qualities of each approach?

1.2 Objectives

To address the research question in the work, the following objectives were derived. Firstly, the dataset will be prepared and aligned to be compatible with the multimodal transformer required input. The second objective is to implement various multimodal transformer harmonization strategies. The third objective will require obtaining the best hyperparameter for each harmonization strategy and training model. Finally, the fourth objective is to evaluate the model using Mean Absolute Error (MAE), Pearson Correlation Coefficient, parameter size, training time and comparison with the dataset benchmark score.

1.3 Contribution of the Work

The major contribution of this work is to help researchers make better decisions in the choice of data fusion techniques for multimodal transformers. Also, it will give insight into the possible requirements such as the model's speed, size, computational cost and expected performance. This will enhance a better choice for multimodal systems applications suitable for human-computer interaction (HCI), telemedicine, customer support and disease diagnosis (Geetha et al., 2024).

1.4 Structure of the Report

The structure of this report goes as follows: Section one is the introduction which comprises the background history, study motivation, research question, objectives, contribution of the work and paper structure. Section two includes several related works on multimodal transformers sectioned based on modality pairs. Section three provides the methodology, focusing on the dataset, transformer model and data fusion strategies. Section four gives the design specification for project work. Section five provides the implementation of the project. Section six contains the evaluation, providing results for experiment 1 (early concatenation fusion), experiment 2 (cross-modal attention fusion), experiment 3 (hierarchical modal attention fusion) and the discussion of the results. Finally, section seven presents the conclusion and future work.

2 Related Work

Several works have been done on multimodal machine learning and transformers using different modality pairs. This section will discuss different multimodal machine learning based on the modality pair. The following modality pairs will be considered, textual and visual modality pair, audio and textual modality pair, audio and visual modality pair and audio, textual and visual modality pair.

2.1 Textual and Visual Modality Pair

Textual data is mostly the based modality for other modalities to complement. (Yao and Wan, 2020) integrates visual information from images with textual context for multimodal machine translation. The multimodal self-attention mechanism within the transformer architecture was used for data fusion with the high interdependency of both modalities. This serves as a significant limitation to this approach as irrelevant or unaligned data from any modality can degrade the performance of the model. The effectiveness and robustness of several attention mechanisms were compared by (Hendricks et al., 2021). The zero-shot task was used to prove the better performance of multimodal attention, even though this might not be robust enough to capture the nuance across all downstream tasks. However, the fusion techniques of these modalities play a crucial role in the robustness and performance of the model. (Chen, Zhang, Li, Deng, Tan, Xu, Huang, Si and Chen, 2022) implemented a multi-level fusion for a hybrid transformer architecture. Integrating the visual and textual data using cross-modal learning supported the claim of the effectiveness of multimodal systems. However, the complexity of the models requires significant computational resources for training and interference.

2.2 Audio and Textual Modality Pair

The audio and textual modalities are both sequential data formats, (Pham et al., 2023) implemented a concatenation approach for a multimodal transformer using audio and textual modalities as input. The BERT was used for textual feature extraction while the audio data was converted to Mel-Spectogram and VGGish was used for feature extraction. The approach yielded a better performance over a single modality. (Chen, Xing, Xu, Yang and Pang, 2022) also proposed the use of a pre-trained model for feature extraction, RoBERTa for text data and Wav2vec for audio data. Cross modality was used for information learning across the modalities after which deep fusion was used to combine information from both modalities. The output was concatenated before generating the output. (Zhang et al., 2023) provide a support claim to the same methodology converting audio file to MFCC and using Bi-LSTM with intra-modal attention for feature extraction while BERT was used for text data. The cross-modal attention was used to share information gained between modalities after which the output representation was concatenated and fed to a fully connected layer to generate an output.

This shows cross-modal attention and concatenation are effective for audio and textual modality fusion. However, the limitation of these works is the dependency on overlapping information from both modalities, both modalities contain the message but with some contextual difference and as a result, improper embedding or excessive noise from any modality will significantly affect the result.

2.3 Audio and Visual Modality Pair

The audio and visual modalities have some inherent challenges such as high computational cost and high level of noise which is minimal in textual data. (Huang et al., 2020) proposed the use of a Transformer with LSTM, the LSTM layer was introduced before the final linear layer in the model. The cross-modal representation was used for modality information fusion. However, this model still poses the challenge of high computational cost attached to both modalities. (Park and Choi, 2024) in light to reduce the computational requirement for multimodal transformer while maintaining the performance of the model using the audio and video data with a proposed Low-Cost Multimodal Transformer (LoCoMT). The model uses predefined attention patterns for each attention head, this is applied to different layers of the Transformer thereby reducing the number of operations required during training. However, to address the challenge of noise in these modalities, (Waligora et al., 2024) proposed a joint multimodal transformer which incorporates joint representation and hierarchical fusion mechanism. This method uses self-attention to compute intra-modality features and cross-modality for inter-modality features, the self-attention features were used to create a joint representation leading to hierarchical fusion techniques. The resulting model is however complex making it so computationally expensive and the performance could vary with a modality with poor data quality.

Furthermore, (John and Kawanishi, 2022) introduced a branch of input called block embedding to the transformer architecture. This is a cross-attention of the audio and video input together. The transformer possesses self-attention for each modality (audio and video) and cross-attention for both modalities. Two different CNN models were used for feature embedding of both modalities after the Log-Mel Spectrogram was used for extracting audio features and MobileNet for video features. The model produced a better result but with a trade-off of increasing the transformer size which in turn increases the computation complexity. Although to reduce the computational cost, 6 frames were

samples from each video, this could lead to the video data containing unaligned data if the sampled frames do not contain relevant information which could reduce the performance of the system. (Geetha et al., 2024) proposed an AuxFormer framework which has three networks: the main audiovisual fusion network, the auxiliary acoustics network and the auxiliary visual network. The audiovisual fusion network has a transformer architecture with the query vectors obtained from the auxiliary networks. Each modality is treated separately with its self-attention mechanism, and a late fusion concatenation was implemented before using a multi-layer perceptron (MLP) to generate a result. The two auxiliary networks are identical with each handling different modalities, the networks also use a multi-head self-attention mechanism. The framework carefully addresses both modalities without over-reliance on any modality. However, the framework could underperform when one modality is noisy or does not provide rich contextual information.

2.4 Audio, Textual and Visual Modality Pair

The audio, textual and visual modalities comprehensively contain wholesome contextual information for communication but not without its limitations of data redundancy and modality information overlap. This brought the need for an adequate approach for extracting information from each modality and fusion strategy. (Shayaninasab and Babaali, 2024) used three pre-trained transformers, BERT for text, Wav2Vec for audio and VideoMAE for video been the best-performing models for each modality out of the selected models. Early fusion and late fusion techniques were used, trying out concatenation and summation of representation vectors with SVM, neural network and XGBoost. The early fusion performs best with the summation of the representation vectors producing the best accuracy of 74.84% on the IEMOCAP dataset. The effectiveness of this model depends on the size and quality of the dataset as multiple pre-trained models will be fine-turned, although the model is less complex and will require less computational cost.

In addition, (Le et al., 2023) proposed a transformer-based fusion using a concatenation of temporal information captured by a transformer encoder for sequential inputs of audio and video data. ALBERT was used for text data preprocessing, while different CNN models were used for audio and video modalities. This work implemented the decoder part of the transformer model using learned embedding and cross-attention to generate output. Although the implementation gave a balance to all modalities the learned embedding utilized by the decoder part could significantly impact the performance of the model. (Siriwardhana et al., 2020) further proposed the use of pre-trained models for feature extraction from the three modalities, audio, text and video data. The Wav2vec was used for audio data, RoBEERTa for text and Fabnet for Video. The inter-modality-attention (IMA) based fusion layer consisting of six transformer blocks was used. The IMA possesses the representation of modality with information gained from other modalities. The output of all the blocks was paired based on the core modality of the block, the paired output for each modality was combined using Hadamard product before concatenating all the results for all modalities. This work employed the use of pre-trained models reducing the training time for each feature extraction model. Meanwhile, the use of six transformer blocks and Hadamard Product before concatenation increased the number of computations and size of the model requiring more computational resources to train. However, the resulting sequence from incorporating audio, text and visual data is always enormous especially if the visual data are video files and this always results in higher computation. As a result, (Sahay et al., 2020) proposed a low-rank fusion for multimodal

transformer using CMU-MOSEI, CMU-MOSI and IEMOCAP datasets. This work implemented a low-rank matrix factorization (LMF) to capture inter-modal signals, this tensor fusion approach models the unimodal, bimodal and trimodal interactions without generating a large multiple representation for each modality embeddings and interaction. The captured signals across all modalities were fused using cross-modal attention. The LMF reduces the size of the model representation and the number of parameters in the model, reducing the model's complexity and making it faster. Although the model performs fairly, it could not outperform the baseline Multimodal Transformer model signifying loss of information during the model minimization.

The related literature shows the strength of multimodal systems with cross-modal attention and concatenation offering the best data fusion strategy, but with a trade-off of requiring more computation. This proves the need for a critical examination of these methods to obtain the best trade-off for these strategies which is the focus of this work.

3 Methodology

3.1 Dataset

The modalities considered for this project are audio, textual and visual. The project was streamlined to emotion recognition downstream task, the CMU-MOSI (Multimodal Corpus of Sentiment Intensity) used for this work can be found on GitHub [1] (Zadeh et al., 2016; Liang et al., 2021, 2023). The dataset consists of extracted audio, text and vision features, the dataset has been annotated with sentiment intensity. The annotation ranges from -3 (strongly negative) to +3 (strongly positive), and the annotation is labelled in continuous values with the specified range. The videos were collected from the YouTube website, the videos have 89 distinct speakers, 41 females and 48 males all from different ethnicity but speaking English. The facial gesture was focused on for the video data, smiles, nods, frowning and head shakes were noted. Several techniques including the Mel-Frequency Cepstral Coefficients (MFCC), COVAREP, and Normalized Amplitude Quotient (NAQ) were used for extracting features from the audio data. The dataset has 2199 video clips generated from 93 videos with a clip average length of 4.2 sec and an average word count of 12. The textual data was generated by manually transcribing the videos. The dataset comes in 3 splits train, validation and test. The train set has 1283 samples, the validation set has 214 samples and the test set has 686 samples (Zadeh et al., 2016). The dataset distribution across different sets is shown in Figure 1, and the label distribution across the various dataset groups is shown in Figure 2.

[1] `https://github.com/pliang279/MultiBench?tab=readme-ov-file`

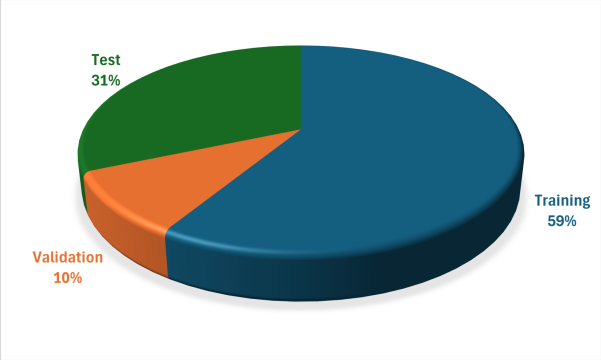

Figure 1: Dataset distribution

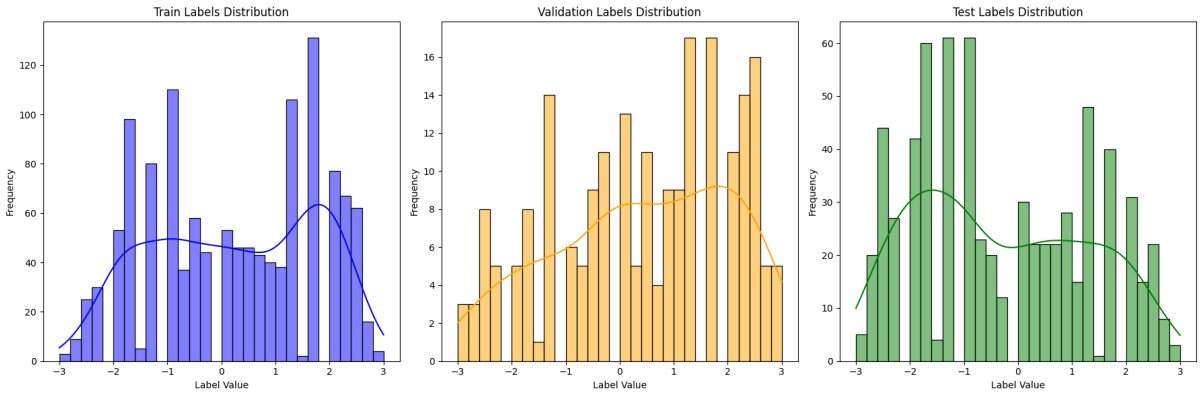

Figure 2: Dataset label distribution

Figure 3 shows the shape of the dataset for all the input modalities across the distributions.

```
mosi_raw.keys()
dict_keys(['train', 'valid', 'test'])

mosi_raw['train'].keys()
dict_keys(['vision', 'audio', 'text', 'labels', 'id'])

print("Vision Data")
print(f"Training Set Shape: {mosi_raw['train']['vision'].shape}")
print(f"Validation Set Shape: {mosi_raw['valid']['vision'].shape}")
print(f"Test Set Shape: {mosi_raw['test']['vision'].shape}")
print("Audio Data")
print(f"Training Set Shape: {mosi_raw['train']['audio'].shape}")
print(f"Validation Set Shape: {mosi_raw['valid']['audio'].shape}")
print(f"Test Set Shape: {mosi_raw['test']['audio'].shape}")
print("Text Data")
print(f"Training Set Shape: {mosi_raw['train']['text'].shape}")
print(f"Validation Set Shape: {mosi_raw['valid']['text'].shape}")
print(f"Test Set Shape: {mosi_raw['test']['text'].shape}")

Vision Data
Training Set Shape: (1283, 50, 35)
Validation Set Shape: (214, 50, 35)
Test Set Shape: (686, 50, 35)
Audio Data
Training Set Shape: (1283, 50, 74)
Validation Set Shape: (214, 50, 74)
Test Set Shape: (686, 50, 74)
Text Data
Training Set Shape: (1283, 50, 300)
Validation Set Shape: (214, 50, 300)
Test Set Shape: (686, 50, 300)
```

Figure 3: Dataset shape

The snippet showing the view of the vision modality is shown in Figure 4.

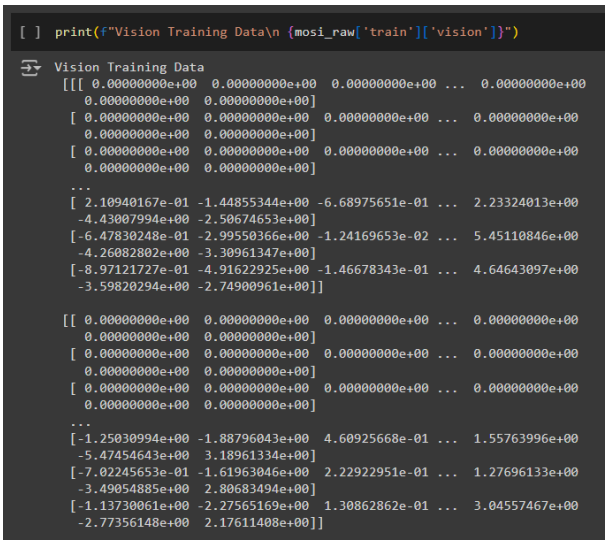

Figure 4: Snippet showing the vision modality data

The snippet showing the view of the textual modality is shown in Figure 5.

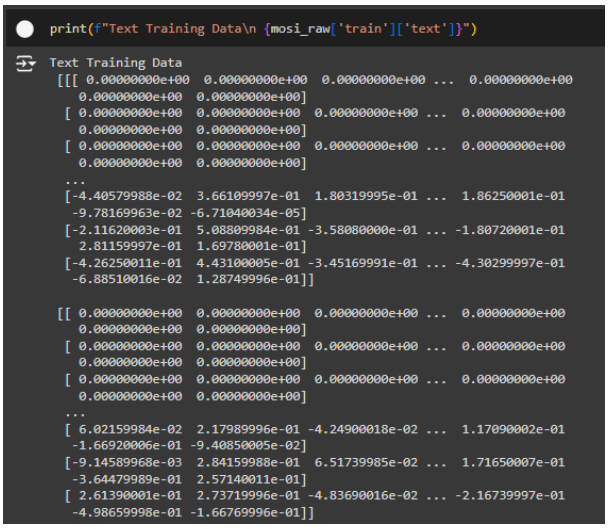

Figure 5: Snippet showing the text modality data

The snippet showing the view of the audio modality is shown in Figure 6.

```
print(f"Audio Training Data\n {mosi_raw['train']['audio']}")

Audio Training Data
 [[[ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   ...
   [ 2.22000000e+02  6.66666687e-01  9.89001915e-02 ...  8.80259797e-02
     4.09209915e-02  1.33800760e-01]
   [ 2.13843750e+02  9.79166687e-01  1.60358369e-01 ... -1.16521502e-02
     6.68889210e-02  1.73782691e-01]
   [ 2.06714279e+02  7.61904776e-01  1.31449461e-01 ...  5.32066263e-02
     1.14056304e-01  2.07019895e-01]]

  [[ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   [ 0.00000000e+00  0.00000000e+00  0.00000000e+00 ...  0.00000000e+00
     0.00000000e+00  0.00000000e+00]
   ...
   [ 2.71062500e+02  2.50000000e-01  2.14687400e-02 ... -8.93318206e-02
    -3.81136909e-02 -5.49064577e-02]
   [ 2.66257568e+02  3.93939406e-01  7.16430545e-02 ... -2.00241953e-01
    -7.68278390e-02 -1.12619251e-02]
   [ 2.50559708e+02  5.37313461e-01  8.56798738e-02 ... -2.01351613e-01
    -1.23198427e-01 -5.77579960e-02]]
```

Figure 6: Snippet showing the audio modality data

3.2 Transformer Model

The base model for this work is the state-of-the-art transformer model titled 'Attention is all you need' (Vaswani et al., 2017). The model has an encoder and a decoder which could be stacked up. The encoder processes the input to extract contextual information which the decoder uses to generate output sequence (Han et al., 2022). The architecture of the transformer model is shown in Figure 7.

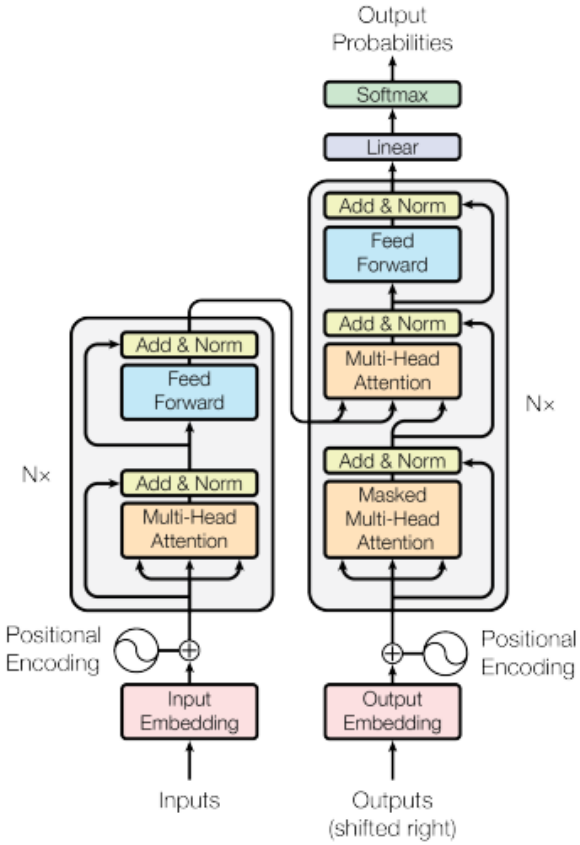

Figure 7: The Transformer model architecture (Vaswani et al., 2017)

The encoder comprises of an input embedding with positional encoding stacked up with encoder layers. The encoder layers are identical blocks consisting of a stack of multi-head attention layers, layer normalization, feed-forward neural network and another layer normalization. The decoder on the other hand consists of an output embedding with positional encoding stacked up with decoder layers. The decoder layers are identical blocks consisting of a masked multi-head attention layer, multi-head attention and feed-forward neural network layers, with each layer succeeded by a layer normalization. The masking in the model ensures the network does not attend to subsequent positions.

The multi-head attention helps to boost the performance by focusing on different positions in the input. The attention function maps the vectors query Q, key K and Value V to generate an output. The formula for calculating self-attention for input vectors is shown in (1), where d_k is the dimension of the keys. The attention function diagram is shown in Figure 8, and the representation of the multi-head attention is given in Figure 9.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

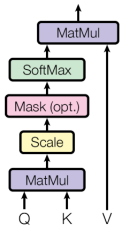

Figure 8: The scale dot-product attention (Vaswani et al., 2017)

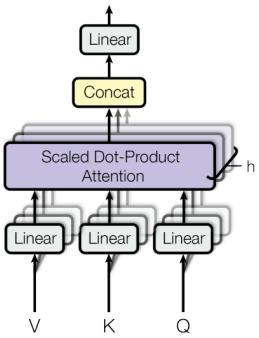

Figure 9: The multi-head attention (Vaswani et al., 2017)

3.3 Fusion Strategy

There are several fusion strategies for multimodal data. However for this low-level data fusion will be considered, some of which exist on the attention-layer. The following fusion strategies are considered early concatenation, cross-modal attention, hierarchical attention and dynamic modal attention.

The early concatenation fusion is a straightforward approach to combining the token embeddings for all modalities as input for the transformer model. This approach always

results in a longer input sequence with no modification to the model than parameter shape and sizes. This method relies on the self-attention of the transformer model. Figure 10 shows the transformer-based early concatenation fusion (Xu et al., 2023).

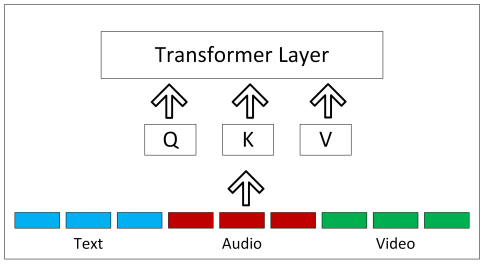

Figure 10: Diagram of early concatenation fusion

However, cross-modal attention fusion relies on the swap of query embedding vectors across the different modalities. This approach maintains the shape and size of the model while sharing information learned across different modalities. This does not cause higher computational complexity compared to concatenation. The illustration for the cross-modal attention fusion is shown in Figure 11 (Xu et al., 2023).

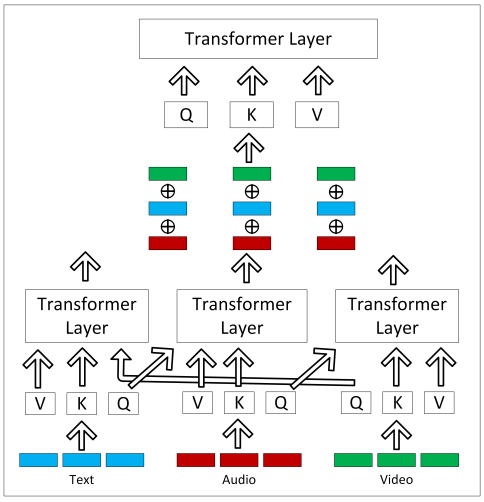

Figure 11: Diagram of cross-modal attention fusion

Furthermore, hierarchical modal attention fusion involves using an independent transformer encoder for each stream of inputs, the outputs are concatenated and fused by another transformer. This approach is a variant of the basic concatenation approach, Figure 12 shows the diagram for hierarchical modal attention fusion (Xu et al., 2023).

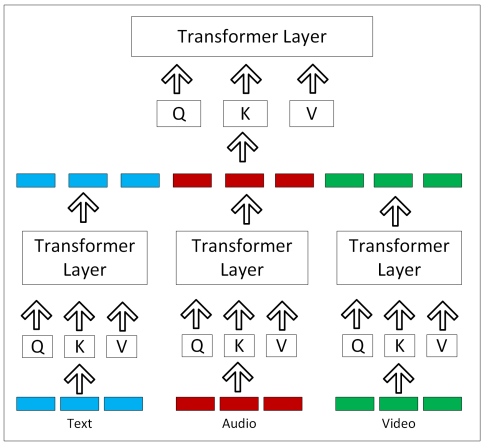

Figure 12: Diagram of hierarchical modal attention fusion

4 Design Specification

The model used for this work was built by stacking the encoder side of the transformer architecture. This is due to the downstream task focused on performance. The model was designed to accept three pre-processed data for text, audio and visual modalities. The transformer models had embedded attention layers for low-level information retrieval. However, the design used for this work does not include the embedding and position encoding layer of the existing models. The exclusion was due to the pre-processing steps and feature extraction performed on the data obtained.

The transformer model used for implementing early concatenation fusion had the input modalities concatenated along the feature dimension axis. Then the data was masked to avoid the model attending to padding in the input data. The resulting output was passed to the transformer encoder with N layer(s), with the resulting output fed into a fully connected layer to generate the sentiment level. The encoder layer in this approach uses a single head attention, this is due to the resulting length of the input sequence after concatenation which is indivisible. Figure 10 shows the representation of the approach.

However, the cross-modal attention mechanism was incorporated on another transformer, the key (K) and Value (V) were retained for each modality while the query (Q) was swapped with another modality. The modalities were paired up resulting in three pairs text-audio, text-video and audio-video enabling information sharing among modalities. Also, each modality has its self-attention as well. The mean of the outputs for the paired modalities' attention and each modality's self-attention was computed. The mean of the outputs was masked, fed into the transformer encoder and then to the fully connected layer. The diagram illustrating the techniques is shown in Figure 11.

In addition, for the hierarchical attention, each modality was fed to a different transformer encoder, and the output was concatenated together and fed to another encoder for modality fusion. The resulting output was fed to a fully connected layer for predictions. The encoder layer utilizes only multi-head self-attention. Figure 12 shows the representation of the approach.

All approaches were connected to a fully connected neural network layer which was connected to an output layer.

5 Implementation

The implementation of this work involved several experiments, hyperparameter search to obtain the best combination for the number of encoder layers and attention heads for each approach other than for the concatenation approach was performed, and a total of 48 trials were run on each method. However, only the best number of encoder layers was obtained for the eaerly concatenation approach, maintaining the number of head to be one. This is due to the indivisible shape dimension obtained after merging all modalities. The resulting number of trials is 12.

The experiment was performed on Google Colab with a core Intel(R) Xeon(R) CPU @ 2.00GHz, RAM of 12.675GB and a single Tesla T4 GPU with 15GB memory. The Adams optimizer with a learning rate of 0.0001. Also, a scheduler was implemented to reduce the learning rate by a factor of 0.8 after no improvement over 3 epochs. The scheduler was implemented to keep the learning rate fit during the training process. In addition, the model was set to train over 100 epochs but the an early stopping to avoid overfitting. The early stopping monitors the validation loss and ends the training after patience of 10 epochs if there is no reduction in the validation loss. The best weights are saved over the training period.

6 Evaluation

The evaluation metrics used to monitor the models' performances are mean absolute error (MAE) and Pearson Correlation. The MAE is a regressive metric measuring the deviation of the predicted value from the actual value. The metrics best suit this task because they provide a generic and bounded performance with no concentration on outliers. The best MAE value is 0 while the worst is $+\infty$ Chicco et al., 2021. A lower MAE indicates a model's predictions are closer to the true values suggesting a better prediction accuracy. The formula for MAE is shown in (2).

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

Where y_i is the target value for ith sample, $\hat{y}_i$ is the predicted value for ith sample and n is the total number of samples.

Pearson's Correlation coefficient is a statistical measure of the linear correlation between two variables. It provides insight into how well the predicted values align with the actual values Sheugh and Alizadeh, 2015. The Pearson's correlation coefficient ranges from -1 to $+1$, where -1 means perfect negative linear correlation, 0 means no linear correlation and $+1$ indicates perfect positive linear correlation. A value closer to $+1$ indicates a strong positive alignment between the predicted and actual values. The formula for Pearson's correlation coefficient is shown on (3).

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \tag{3}$$

where:

- r is the Pearson correlation coefficient,

- y_i is the target value for the ith sample,
- $\hat{y}_i$ is the predicted value for the ith sample,
- $\bar{y}$ is the mean of the target values,
- $\bar{\hat{y}}$ is the mean of the predicted values,
- n is the total number of samples.

In addition to the MAE and Pearson correlation metrics, the size of the resulting model is given by the number of parameters in the model, the training time and equivalent CO_2 emission for the training process.

6.1 Experiment 1: Early Concatenation Fusion

The best 5 results for experiment 1 using early concatenation are shown in Table 1. The plot for the hyperparameters with the best MAE and Correlation score are shown in Figure 13 and Figure 14 respectively.

Table 1: Result for Early Concatenation Fusion.

Number of Encoder Layers	MAE	Correlation Score	Training Time	Parameter Size
2	0.0186	0.4858	16.19 Secs	1,591,720
3	0.0169	**0.4903**	27.93 Secs	2,387,375
5	0.0179	0.4356	29.46 Secs	3,978,685
7	**0.0153**	0.4222	44.50 Secs	5,569,995
9	0.0191	0.4265	51.71 Secs	7,161,305

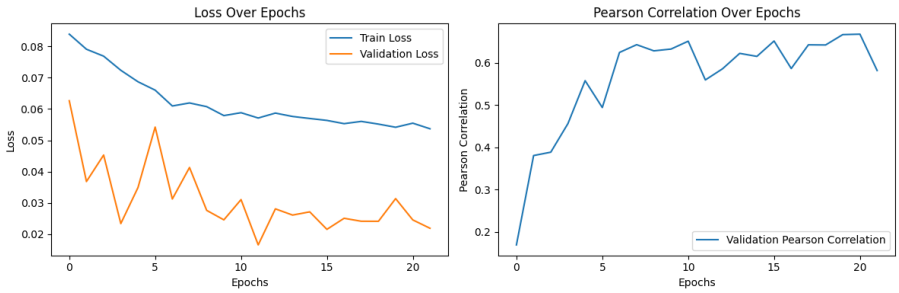

Figure 13: Plot of MAE and Pearson's correlation coefficient for seven encoder layers

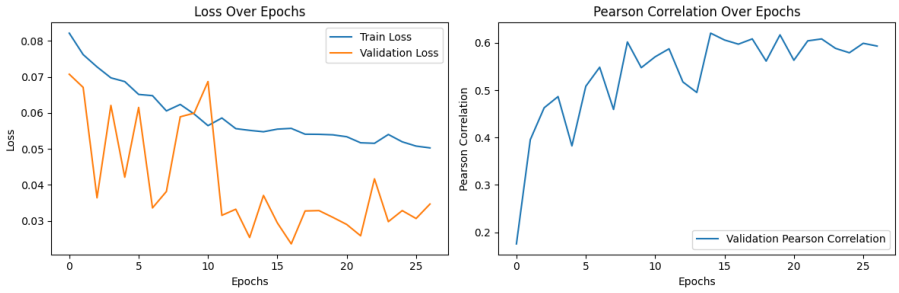

Figure 14: Plot of MAE and Pearson's correlation coefficient for three encoder layers

6.2 Experiment 2: Cross-Modal Attention Fusion

The best 5 results for experiment 2 using cross-modal attention fusion are shown in Table 2. The plot for the hyperparameters with the best MAE and Correlation score are shown in Figure 15 and Figure 16 respectively.

Table 2: Result for Cross-Modal Attention Fusion.

Number of Encoder Layers	Number of Attention Head	MAE	Correlation Score	Training Time	Parameter Size
3	4	0.0155	**0.3956**	64.25 Secs	2,248,387
8	4	0.0168	0.2781	93.72 Secs	3,955,377
8	8	**0.0152**	0.3541	88.08 Secs	3,955,377
8	16	0.0192	0.1941	119.30 Secs	3,955,377
8	32	0.0159	0.2690	121.46 Secs	3,955,377

Figure 15: Plot of MAE and Pearson's correlation coefficient for eight encoder layers and eight multi-head attention

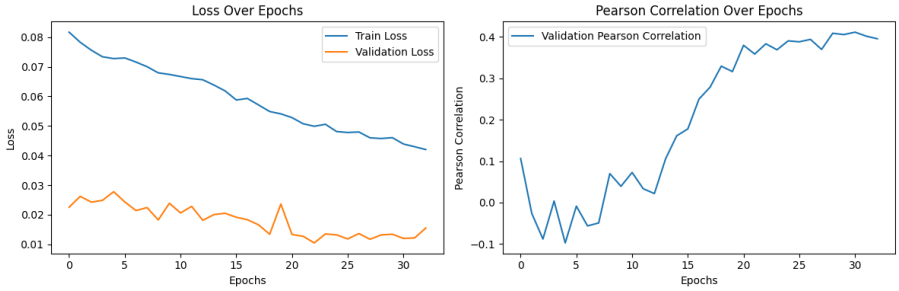

Figure 16: Plot of MAE and Pearson's correlation coefficient for three encoder layers and four multi-head attention

6.3 Experiment 3: Hierarchical Modal Attention Fusion

The best 5 results for experiment 3 using hierarchical modal attention fusion are shown in Table 3. The plot for the hyperparameters with the best MAE and Correlation score is shown in Figure 17 and Figure 18 respectively.

Table 3: Result for Hierarchical Modal Attention Fusion.

Number of Encoder Layers	Number of Attention Head	MAE	Correlation Score	Training Time	Parameter Size
1	2	0.0144	0.5267	42.12 Secs	9,761,793
1	4	**0.0105**	0.5257	44.57 Secs	9,761,793
2	1	0.0111	**0.5509**	54.85 Secs	19,220,993
2	2	0.0146	0.4892	78.22 Secs	19,220,993
4	4	0.0139	0.5133	248.39 Secs	38,139,393

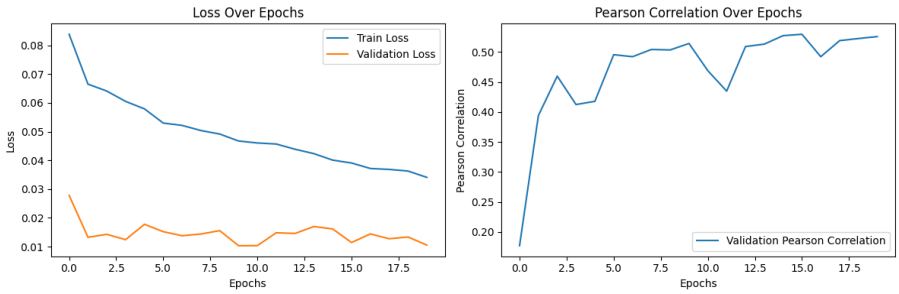

Figure 17: Plot of MAE and Pearson's correlation coefficient for one encoder layer and four multi-head attention

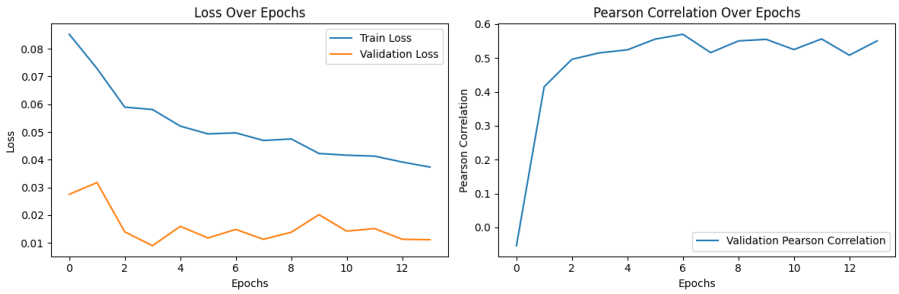

Figure 18: Plot of MAE and Pearson's correlation coefficient for two encoder layers and one multi-head attention

6.4 Discussion

The result obtained in experiment 1, early concatenation fusion strategy shows varying performance but with the lowest MAE (0.0153) achieved using 7 encoder layers. However, the model also has the lowest correlation score (0.4222) in the reported samples indicating that the model could perform well with some samples but could not generalize overall samples. Although the mean absolute error (MAE) is low, indicating a low deviation in the predicted value, the fair correlation score indicates a weak alignment with the true sentiment labels. This suggests limitations in capturing sentiment intensity effectively indicating the fusion strategy is not robust enough.

However, for experiment 2, the cross-modal attention fusion had the best performance using 8 encoders with different numbers of heads. Using 8 encoder layers and 8 attention heads, the lowest MAE of 0.0152 was obtained with a corresponding correlation score of

0.3541. The low correlation score indicates a poor alignment with sentiment predictions compared to early concatenation fusion. Although a better MAE score was obtained, the corresponding correlation highlights a notable disparity between the predicted and actual sentiment intensities.

Furthermore, in experiment 3 the hierarchical modal attention obtained the lowest MAE of 0.0105 with 1 encoder layer and 4 attention heads with a good correlation score of 0.5257. The best correlation score of 0.5509 was obtained using 2 encoder layers and 1 attention head with a relatively low MAE score of 0.0111. The correlation score indicates a strong alignment between the predicted and actual sentiment intensities, with a low MAE indicating a low deviation from the actual sentiment intensities. This fusion approach was able to capture nuances of sentiment intensity across modalities and effectively fuse it for better performance.

Conversely, the hierarchical modal attention model is bigger than the other models, a hierarchical modal attention model with 2 encoder layers has approximately 19 million parameters while the concatenation approach model and cross-modal attention model possess approximately 2 million parameters each for a model with the same number of encoders. This provides an insight into the reason for a better performance. However, the size of concatenation models grows marginally with an increasing number of encoders with the 7-layer encoder layer possessing approximately 6 million parameters while cross-modal attention models with 8 encoder layer possessing approximately 4 million parameters. Although the cross-modal attention model has a lower MAE value, the correlation score is relatively lower than the concatenation approach models.

Also, the cross-modal attention models had longer training time compared to other models signifying a slower convergence rate and more computation.

The comparison between various experiments with the baseline performance of the dataset is shown in Table 4. The hierarchical model attention fusion significantly outperforms the multimodal dictionary (Zadeh et al., 2016) which had an MAE of 1.1 and a correlation score of 0.53.

Table 4: Comparison of Result with Baseline Performance (Zadeh et al., 2016).

Approach	**MAE**	**Correlation Score**
Multimodal Dictionary (Zadeh et al., 2016)	1.100	0.53
Multimodal Transformer (Concatenation)	0.015	0.42
Multimodal Transformer (Cross-Modal Attention)	0.015	0.35
Multimodal Transformer (Hierarchical Modal Attention)	**0.011**	**0.55**

7 Conclusion and Future Work

This project focuses on examining the effectiveness and impact of multimodal transformer data fusion strategies. To achieve this, three multimodal transformer models were built with different data fusion approaches. The early concatenation approach, cross-modal

attention modal and hierarchical modal approach were implemented. The MAE, Pearson correlation coefficient, training time and parameter size were measured for each approach. The hierarchical attention modal produced the best MAE of 0.0111 and the best correlation score of 0.5509. The approach also had the largest parameter size and longest training time. The cross-modal approach had the smallest parameter size with a similar MAE to the early concatenation approach but a poorer correlation score. The early concatenation had the fastest training time even though possessing more parameters than cross-modal attention. The research also showed that models with larger parameter sizes are more robust than smaller ones, but that may not necessarily affect the speed of the model.

However, this work is only performed on a single downstream task, as a result, future work will focus on more downstream tasks to provide more insight into multimodal transformer data fusion strategies. Also, implementing more data fusion strategies for multimodal transformers.

References

Baevski, A., Zhou, Y., Mohamed, A. and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems* **33**: 12449–12460.

Chen, W., Xing, X., Xu, X., Yang, J. and Pang, J. (2022). Key-sparse transformer for multimodal speech emotion recognition, *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6897–6901.

Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L. and Chen, H. (2022). Hybrid transformer with multi-level fusion for multimodal knowledge graph completion, *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 904–915.

Chicco, D., Warrens, M. J. and Jurman, G. (2021). The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, *Peerj computer science* **7**: e623.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* .

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale, *arXiv preprint arXiv:2010.11929* .

Flores-Coronado, M. A., Ciria, A. and Lara, B. (2022). Multimodal integration as predictions. an explanation of the mcgurk effect., *CogSci*.

Geetha, A., Mala, T., Priyanka, D. and Uma, E. (2024). Multimodal emotion recognition with deep learning: advancements, challenges, and future directions, *Information Fusion* **105**: 102218.

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y. et al. (2022). A survey on vision transformer, *IEEE transactions on pattern analysis and machine intelligence* **45**(1): 87–110.

Hendricks, L. A., Mellor, J., Schneider, R., Alayrac, J.-B. and Nematzadeh, A. (2021). Decoupling the role of data, attention, and losses in multimodal transformers, *Transactions of the Association for Computational Linguistics* **9**: 570–585.

Huang, J., Tao, J., Liu, B., Lian, Z. and Niu, M. (2020). Multimodal transformer fusion for continuous emotion recognition, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 3507–3511.

John, V. and Kawanishi, Y. (2022). Audio and video-based emotion recognition using multimodal transformers, *2022 26th International Conference on Pattern Recognition (ICPR)*, IEEE, pp. 2582–2588.

Karani, R. and Desai, S. (2022). Review on multimodal fusion techniques for human emotion recognition, *Int. J. Adv. Comput. Sci. Appl* **13**: 287–296.

Le, H.-D., Lee, G.-S., Kim, S.-H., Kim, S. and Yang, H.-J. (2023). Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning, *IEEE Access* **11**: 14742–14751.

Li, L., Zhang, D., Zhu, S., Li, S. and Zhou, G. (2024). Response generation in multimodal dialogues with split pre-generation and cross-modal contrasting, *Information Processing & Management* **61**(1): 103581.

Liang, P. P., Lyu, Y., Fan, X., Agarwal, A., Cheng, Y., Morency, L.-P. and Salakhutdinov, R. (2023). Multizoo & multibench: A standardized toolkit for multimodal deep learning, *Journal of Machine Learning Research* **24**: 1–7.

Liang, P. P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L. Y., Wu, P., Lee, M. A., Zhu, Y. et al. (2021). Multibench: Multiscale benchmarks for multimodal representation learning, *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* .

Park, S. and Choi, E. (2024). Multimodal transformer with a low-computational-cost guarantee, *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 6125–6129.

Pham, N. T., Dang, D. N. M., Pham, B. N. H. and Nguyen, S. D. (2023). Server: Multimodal speech emotion recognition using transformer-based and vision-based embeddings, *Proceedings of the 2023 8th International Conference on Intelligent Information Technology*, pp. 234–238.

Sahay, S., Okur, E., Kumar, S. H. and Nachman, L. (2020). Low rank fusion based transformers for multimodal sequences, *arXiv preprint arXiv:2007.02038* .

Shayaninasab, M. and Babaali, B. (2024). Multi-modal emotion recognition by text, speech and video using pretrained transformers, *arXiv preprint arXiv:2402.07327* .

Sheugh, L. and Alizadeh, S. H. (2015). A note on pearson correlation coefficient as a metric of similarity in recommender system, *2015 AI & Robotics (IRANOPEN)*, IEEE, pp. 1–6.

Siriwardhana, S., Kaluarachchi, T., Billinghurst, M. and Nanayakkara, S. (2020). Multimodal emotion recognition with transformer-based self supervised feature fusion, *Ieee Access* **8**: 176274–176285.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F. and Dai, J. (2019). Vl-bert: Pre-training of generic visual-linguistic representations, *arXiv preprint arXiv:1908.08530* .

Taheri, Z. S., Roy, A. C. and Kabir, A. (2023). Bemofusionnet: A deep learning approach for multimodal emotion classification in bangla social media posts, *2023 26th International Conference on Computer and Information Technology (ICCIT)*, IEEE, pp. 1–6.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need, *Advances in neural information processing systems* **30**.

Waligora, P., Aslam, M. H., Zeeshan, M. O., Belharbi, S., Koerich, A. L., Pedersoli, M., Bacon, S. and Granger, E. (2024). Joint multimodal transformer for emotion recognition in the wild, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4625–4635.

Xu, P., Zhu, X. and Clifton, D. A. (2023). Multimodal learning with transformers: A survey, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **45**(10): 12113–12132.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding, *Advances in neural information processing systems* **32**.

Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation, *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4346–4350.

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 preprint arXiv:1606.06259* .

Zhang, S., Yang, Y., Chen, C., Liu, R., Tao, X., Guo, W., Xu, Y. and Zhao, X. (2023). Multimodal emotion recognition based on audio and text by using hybrid attention networks, *Biomedical Signal Processing and Control* **85**: 105052.