Biprojection Multimodal Transformer for Multimodal Data Classification[⋆,⋆⋆]

Diego Aarón Moreno-Galván[a,*] (Student), Adrián Pastor López-Monroy[a,**] (Assessor) and Luis Carlos González-Gurrola[b,*] (Co-assessor)

[a]Centro de Investigación en Matemáticas (CIMAT), Jalisco S/N, Guanajuato, 36262 GT, México
[b]Universidad Autónoma de Chihuahua (UACH), Chihuahua, 695014, Chihuahua, México

ARTICLE INFO

Keywords:
Multimodal Classification
Transformers
Movie Genre Classification
Emotion Detection
Multimodal Transformer
BERT
GMU
Multimodal Fusion
Deep Learning

ABSTRACT

Analyzing, manipulating, and comprehending data from multiple sources (e.g., websites, software applications, files, or databases) has become increasingly important. For instance, many digital platforms, such as Netflix or YouTube, are interested in recommending appropriate and relevant digital material for us based on multimodal information, such as language turned into text, images, and audio. Current research centers on developing efficient strategies to automatically analyze and comprehend this type of multimodal content for classification. Furthermore, multimodal data is also used to detect emotions and sentiments in behavioral investigations. The movie genre categorization task is an exciting research case among the numerous problems in the modality domain. We take it in this work as a study case with two distinct datasets. Moreover, we also take the emotion recognition task as a study case and conduct several experiments with another two datasets.

A recently employed technique for multimodal categorization is the Multimodal Transformer (MulT) model, which is based on taking relevant information using attention matrices of each modality. Our research primarily addresses the automatic and efficient modality combining issue within the MulT model. The MulT has an excellent performance on many supervised multimodal datasets, but the information from various modalities is fused with a heavy transformer and is not intuitive. Furthermore, we focus on leveraging the GMU module within the architecture to efficiently and dynamically weigh modalities at the instance level and to comprehend and visualize the use of modalities. Moreover, a common challenge in deep learning is when the model fails to learn due to vanishing gradients; to overcome this issue, we focus on strategically placing residual connections in the architecture. We propose a novel architecture to compete with current state-of-the-art (SOTA) models in movie genre classification, outperforming them by 2% on Moviescope and 1% on MM-IMDB datasets. In the emotion recognition task, we get competitive performance with SOTA models on the IEMOCAP dataset, and we improve by 1% on the CMU-MOSEI[1] dataset.

1. Introduction

Multimodal classification has become increasingly relevant for analyzing data from many sources, Baltrušaitis, Ahuja and Morency (2017), and Xu, Zhu and Clifton (2022). For instance, many digital platforms are interested in recommending relevant and appropriate content for us. Besides, multimodal data are also present in detecting emotions and sentiments for behavior studies, for example, the Social Anxiety Disorder Nikolić, Majdandžić, Colonnesi, de Vente, Möller and Bögels (2020), the engagement while learning Charland, Léger, Sénécal and Courtemanche (2015), or the stress detection Yao, Papakostas, Burzo, Abouelenien and Mihalcea (2021). The word "multimodal" refers to using the information from various representations or transmission channels to perform classification, e.g., in a video, we can obtain the language which can be transformed into text, the sequence of images, and the audio. The effective use of these modalities is a current area of research called multimodal classification task, Sleeman-IV, Kapoor and Ghosh (2021).

In our case, we focus on the movie genre classification task and perform different experiments on the multimodal emotion recognition task. We use the movie genre categorization task because it is a multimodal problem (which may involve video, audio, and text). Also, there is a recent open-access multimodal movie genre classification dataset called Moviescope. Moreover, this task is essential in our lives to have appropriate content when using social media or to see a movie. On the other hand, we use the emotion recognition task because it also is a multimodal problem since we can analyze gestures, intonation, and the words that a person says to detect their sentiment. This task is commonly used to compare models that work with multimodal data, and we introduce our proposed model into this benchmark. Traditional text-based approaches to sentiment analysis collect large amounts of text data, from which various algorithms are used to extract sentiment information. However, multimodal sentiment analysis offers a way to perform opinion

⋆This document is the results of the author's research thesis to obtain his Master degree funded by the CIMAT.

⋆⋆Biprojection Multimodal Trasformer is a novel architecture based on the original Multimodal Transformer reordering the Crossmodal Transformer blocks to get a better recognition.

*Corresponding author
**Principal corresponding author
diego.moreno@cimat.mx (D.A. Moreno-Galván); pastor.lopez@cimat.mx (A.P. López-Monroy); lcgonzalez@uach.mx (L.C. González-Gurrola)
damorgal.io (D.A. Moreno-Galván); www.cimat.mx/pastor (A.P. López-Monroy); www.uach.mx/lcgonzalez (L.C. González-Gurrola)
ORCID(s): 0000-0002-8433-9591 (D.A. Moreno-Galván); 0000-000x-xxxx-xxxx (A.P. López-Monroy); 0000-000x-xxxx-xxxx (L.C. González-Gurrola)

analysis based on a combination of video, audio, and text that goes far beyond traditional text-based sentiment analysis in understanding human behavior, Chauhan, Sharma and Sikka (2021). Furthermore, this task finds various applications like movie reviews on YouTube, product opinions, or health care applications, including stress and depression analysis, Chandrasekaran, Nguyen and Hemanth (2021).

In this context of multimodal data analysis, various methods were designed to handle and combine data from many sources. A typical recent model mainly used in supervised classification is a transformer-based neural network called Multimodal Transformer (**MulT**). Some recent works for multimodal classification use the MulT model as their base form. The problem with this model is that it fusions the information from a distinct source with a Transformer which is a heavy architecture, and it is not clear what modalities are taken into account for the final decision. Our research focuses primarily on solving modalities combination within the MulT model, adding a dynamic combination module and a reorder in the crossmodal transformers called **biprojections** in this work. We proposed a novel architecture to tackle current state-of-the-art models in the movie genre classification, the Biprojection Multimodal Transformer (**BPMulT**). We extended our experiments to the emotion recognition task to introduce our proposed model into this standard used multimodal benchmark.

Therefore the main contributions of this thesis are fivefold:

1. This BPMulT considers a reorder of crossmodal transformers called biprojection to enrich each modality with the other sequences. It is a crucial step because this module reordering allows a better flow of information, introduces residual connections, and finds better patterns across the modalities—our second contribution related to the combination of multimodal information.
2. We introduce the Fusion Gated Multimodal Unit (FGMU) based on the GMU, Arevalo, Solorio, Montes-y Gómez and González (2017) to fuse information from the same modality. This module modifies the original GMU, just adding a residual product inside the module. It is an essential piece because we show that if we use the original GMU, it affects the classification performance.
3. Our third contribution is introducing two types of residual connections to address the efficient learning problem in deep neural networks. One type is an adding connection between the biprojections, and the second is a fusion of projections followed by an adding. We showed that it is the best configuration for better classification performance.
4. The fourth contribution is related to a novel taxonomy inspired by the hybrid method. We called this taxonomy a parallel method where one part learns to classify heavily, and the other learns lightly. To reduce the over-fitting, we add to our model a parallel and simple modalities fusion to find patterns in sequences at a high level. Then, we fuse these parallel architectures (light and heavy) with a GMU to dynamically fuse and predict.
5. Finally, our fifth contribution is that we outperform the proposed model's current state-of-the-art results in Moviescope and MM-IMDb for movie genre classification. It has competitive performance in the emotion recognition task.

2. Related Work

We show a selection of previous publications that has the most relevant relation to our research regarding datasets, models, and components.

2.1. MM-IMDb and GMU

The work of Arevalo et al. (2017) was one of the first publications for movie genre classification tasks. This work produced the Multimodal IMDb (MM-IMDb) dataset for this task. They also proposed a novel module for the dynamic combination of two modalities called the Gated Multimodal Unit (GMU). It is a crucial publication for our research since we perform experiments in the MM-IMDb dataset, we use the GMU in our proposed model to fuse multimodal information, and we propose a variation of this module with a residual connection inside the architecture.

Their primary motivation was that existing similar datasets for multimodal classification, at that time, were small, containing less than 10,000 movies. Additionally, they integrate extra and helpful information from the IMDb website for a better understanding, like plot, poster, and metadata.

The principal contributions are the multimodal MM-IMDb dataset and the Gated Multimodal Unit (GMU). For our experiments, we use the MM-IMDb for comparison with this publication and the MMBT model. The GMU is a module compounded by gates that allow dynamically classifying based on several modalities' relevance. We are using this module as a part of our final architecture.

2.2. Multimodal BiTransformer

This work Kiela, Bhooshan, Firooz, Perez and Testuggine (2019) uses the MM-IMDb dataset to compare their proposed model (MMBT) against GMU. They perform better on this dataset considering just the poster and text movie data.

The inspiration of the MMBT model is to take advantage of the pre-trained BERT model (on text modality) by embedding the visual features into the text token space. The model learns how to integrate and fuse embedded visual features with the text features and then performs the multimodal classification. Therefore, they perform fine-tuning like a regular BERT model.

Image features are extracted from a pre-trained ResNet-152, He, Zhang, Ren and Sun (2015). They extract N features because the BERT model handles sequences, so they define the N features of an image as a sequence of length N. Hence, they extract N features of dimension 2048 from a single image. Then, they perform a pooling operation instead of

taking the final fully connected layer of the ResNet-152. They project the image features to dimension $D = 768$ the BERT's hidden dimension with:

$$I_n = W_n f(x, n) \in \mathbb{R}^D, W_n \in R^{2048 \times D}, \quad (1)$$

where $f(x, n)$ is the nth visual feature obtained from the ResNet-152 followed by a pooling operation. The final visual representation is obtained by adding each projected feature I_n with the corresponding position and segment token, like in the BERT model. Fine-tuning and prediction are similar to the BERT model with the CLS token for prediction.

As we mentioned, this model was trained on three different sets, and one of them was the MM-IMDB dataset. MMBT outperforms the GMU model, and we will compare our proposed model with these classification results.

2.3. Multimodal Transformer with GMU (MulT-GMU)

The Rodríguez-Bribiesca, López-Monroy and y Gómez (2021) work is our primary base model and inspiration for our research. The authors propose to classify movie genres of the Moviescope dataset using three main components: the pre-trained BERT model to represent word sequences, the Multimodal Transformer to fuse the text, video, and audio features, and a GMU module to combine each modality and give a prediction dynamically. For audio, video, and metadata features, they use the given pre-processed features in Cascante-Bonilla, Sitaraman, Luo and Ordonez (2019).

The authors present their proposed MulT-GMU as an extension of the MulT model combined with the GMU module to boost its capacity. GMU weighs modalities' relevance and adds interpretability.

They perform experiments with their proposed model using different modalities. For example: text + video + audio against text + video. In conclusion, they found that the best performance is when the modalities combination is compounded by text, video, audio, and poster. Metadata adds complexity to the model, and they need to avoid essential transformer attention for better results. Hence, for our experiments, we use just the text, video, audio, and poster features for a fair comparison.

2.4. Multimodal End-to-End for Emotion Recognition

In this work Dai et al. (2021), the author solves the existing problem on multimodal affective computing tasks, such as emotion recognition. The problem is that, generally, models use a two-phase pipeline. The first phase is to extract feature representations for each modality with hand-crafted algorithms. The second is to perform end-to-end learning with the extracted features. Note that extracted features are fixed and cannot be further fine-tuned, which does not generalize or scale well to different tasks. To solve this, the authors develop a fully end-to-end model that jointly connects the two phases and optimizes them. Hence, they restructure the CMU-MOSEI dataset (and others) to enable fully end-to-end training. Also, they introduce a sparse cross-modal attention mechanism for the feature extraction, and their fully end-to-end model is the current state-of-the-art model.

We are using this modified version of the CMU-MOSEI[1] dataset to perform our comparison experiments since this model is the current state-of-the-art.

3. Problems in the MulT

We detect specific problems within the MulT architecture and we focus our research on objectives related to this issues.

1) The modalities interaction inside the MulT model, Figure 4, is given by a complete search of each modality pair combination using the Crossmodal Transformer. We say in this work that a modalities combination is a projection from one modality space to another. For example, in the Yellow Crossmodal Transformer blocks of Figure 4, there is a projection from Video (V) to the Language (L) (V—→L). In this context, a simple projection from the space of modality B to the space of A does not rescue any information of other different modality C because B does not have access to C. In the example, the projection from Video to Text (language) does not have any information on the Audio modality. We hypothesize that it is a problem in the MulT that we can solve by taking a **biprojection**. A Biprojection first involves projecting modality C to the space of B, then B would have relevant information of C and B. Therefore, we project the modified modality B to the space of A to have an enriched modality representation. In the example, we first take a projection from Audio to Video, then this Video features projected to the Text modality.

2) Secondly, the MulT model combines two representations of one modality by a transformer. For example, it takes two simple projections B to A and C to A to get an enriched representation of modality A. In Figure 4, we can see what we mentioned after the Crossmodal Transformers concatenation. The transformer to combine these vectors takes a lot of memory space and is hard to interpret. We believe that we can solve this problem by substituting this transformer for a proposed dynamic fusion module based on the GMU architecture, Arevalo et al. (2017).

3) Finally, the inefficient learning caused by the vanishing gradient is a prevalent problem in deep learning neural networks. Note that the MulT model has just residual connections inside the Crossmodal Transformer but does not have any outside these modules to address this problem. We hypothesize that adding strategical residual connections between levels of these projections will help the model to learn.

[1] http://immortal.multicomp.cs.cmu.edu/raw_datasets/processed_data/cmu-mosei/seq_length_50/mosei_senti_data_noalign.pkl

4. Proposed Approach

This section describes the proposed architecture called **Biprojection Multimodal Transformer (BPMulT)** that aims to solve the problems discussed in Section 3. We observe the proposed Biprojection Multimodal Transformer (BPMulT) architecture in Figure 4.

For the experiments, we also use a lighter version of our model to compare performance when we do not care about over-fitting. We will see that the BPMulT architecture contains a simple parallel part. We call the lighter version BPMulT-no-parallel which has no parallel part. Indeed, the BPMulT-no-parallel does not have the modalities' general and straightforward information, so it over-fits the data.

Our proposed model is compounded by five main parts described below and divided into subsections. The first module (4.1) corresponds to the low-level Crossmodal Transformers, called for us: projections that are explained with our notation to have clarity, but they are precisely the same in the MulT model, Tsai, Bai, Liang, Kolter, Morency and Salakhutdinov (2019). The second one (4.2) is a second Crossmodal Transformer, which is the proposed biprojection and is the solution to our first detected problem. The following section (4.4) is a proposed GMU module for a dynamic and interpretable fusion of modalities called Fusion GMU (FGMU). The FGMU corresponds to the solution of our second detected problem. The fourth module (4.5) corresponds to the parallel fusion of modality vectors we propose not to over-fit the data. The final part of this chapter (4.9) describes the strategic components of our proposed architecture used to solve vanishing gradient problems. It is the solution to our third detected problem.

As an input, our model will consider three primary modalities: the text modality or language (L), vision modality with video frames (V), and audio (A) from a given multimodal dataset. For each modality, we have a feature sequence denoted as $X_M \in \mathbb{R}^{S_M \times d_M}$, where $M \in \{L, V, A\}$, and S_M and d_M corresponds to sequence length and feature dimension of M modality, respectively.

4.1. First Crossmodal Projections

Based on Tsai et al. (2019), the MulT model uses crossmodal projections to find relevant patterns of modality M_1 in modality M_2 space ($M_1 \rightarrow M_2$), where $M_1 \neq M_2$, and $M_1, M_2 \in \{L, V, A\}$. We are using exactly these kinds of projections as our first module. Following the MulT this module is compounded by the Crossmodal Transformer layers with the same Crossmodal Attention, Temporal Convolutions, and Positional Encoding.

4.2. Crossmodal Biprojection

This section is one of the most important in the proposal because it constitutes the primary motivation (the biprojection).

Taking the six possible Crossmodal Transformers (CT) between (L, V, A) modalities, second crossmodal attention is proposed to ensure that the adaptation to a modality M_3 contains information of the other set of two modalities M_1 and M_2.

We are using the CT output sequences of $(M_1 \rightarrow M_2)$ and $(M_2 \rightarrow M_1)$ in order to perform a third CT $(M_1 \rightarrow M_2 \rightrightarrows M_3)$ or $(M_2 \rightarrow M_1 \rightrightarrows M_3)$.

4.3. Second Crossmodal Transformers

Similar to the first level of crossmodal transformers, we use the exact attention mechanism described before.

Following the last section notation, let be $\hat{X}_{M_3} \in \mathbb{R}^{S_{M_3} \times d}$ the input feature sequence defined in Equation (**??**), and $Z^{[D]}_{M_1 \rightarrow M_2} \in \mathbb{R}^{S_{M_2} \times d}$ the last hidden state of a CT. Hence, crossmodal attention for M_3 is given by $Y^{(1,2)}_{M_3} \in \mathbb{R}^{S_{M_3} \times d}$ as follows:

$$Y^{(1)}_{M_3} = \underset{M_1 \rightarrow M_2 \rightrightarrows M_3}{\mathbf{CM}}(\hat{X}_{M_3}, Z^{[D]}_{M_1 \rightarrow M_2}) \tag{2}$$

$$= \text{softmax}\left(\frac{Q_{M_3} K^T_{M_1 \rightarrow M_2}}{\sqrt{d}}\right) V_{M_1 \rightarrow M_2} \tag{3}$$

$$= \text{softmax}\left(\frac{\hat{X}_{M_3} W_{Q_{M_3}} W^T_{K_{M_1 \rightarrow M_2}} Z^{[D]\,T}_{M_1 \rightarrow M_2}}{\sqrt{d}}\right) Z^{[D]}_{M_1 \rightarrow M_2} W_{V_{M_1 \rightarrow M_2}}, \tag{4}$$

and analogous, $Y^{(2)}_{M_3}$ is defined just permuting the order of the first projections M_1 and M_2 as:

$$Y^{(2)}_{M_3} = \underset{M_2 \rightarrow M_1 \rightrightarrows M_3}{\mathbf{CM}}(\hat{X}_{M_3}, Z^{[D]}_{M_2 \rightarrow M_1}), \tag{5}$$

where $W_{Q_{M_3}}, W_{K_{M_1 \rightarrow M_2}}, W_{V_{M_1 \rightarrow M_2}} \in \mathbb{R}^{d \times d}$ are learnable weights.

Note that in this case, the (i, j)th entry of the softmax part is measuring the attention given by the ith time step (or token) of modality M_3 to the jth time step of modality M_2 which at the same time, the jth time step of M_2 is a weighted summary of V_{M_1}. It is the main reason to interpret the second crossmodal projection to modality M_3 as a weighted summary of the other modalities M_1 and M_2.

Then, we complete the Crossmodal Transformer (CT) architecture stacking again $D = 3$, CT layers as described below:

$$Z^{[0]}_{M_1 \rightarrow M_2 \rightrightarrows M_3} = Z^{[0]}_{M_3} = \hat{X}_{M_3}, \tag{6}$$

$$\hat{Z}^{[i]}_{M_1 \rightarrow M_2 \rightrightarrows M_3} = \underset{M_1 \rightarrow M_2 \rightrightarrows M_3}{\mathbf{CM}}\left[\mathbf{LN}\left(Z^{[i-1]}_{M_1 \rightarrow M_2 \rightrightarrows M_3}\right), \mathbf{LN}\left(Z^{[D]}_{M_1 \rightarrow M_2}\right)\right] + \mathbf{LN}\left(Z^{[i-1]}_{M_1 \rightarrow M_2 \rightrightarrows M_3}\right), \tag{7}$$

$$Z^{[i]}_{M_1 \rightarrow M_2 \rightrightarrows M_3} = f^{[i]}_{\theta_{M_1 \rightarrow M_2 \rightrightarrows M_3}}\left[\mathbf{LN}\left(\hat{Z}^{[i]}_{M_1 \rightarrow M_2 \rightrightarrows M_3}\right)\right] + \mathbf{LN}\left(\hat{Z}^{[i]}_{M_1 \rightarrow M_2 \rightrightarrows M_3}\right) \tag{8}$$

Once we have the CT's last output sequence, we obtain a single modality vector for classification. Note that at this time, for each modality, we have two sequence representations. In the following subsection, there is an explanation of how these two modality sequences are fused.

4.4. Modalities Fusion with FGMU

This section is also one of the most critical parts of the proposal because it contains the proposed fusion with the Fusion GMU (FGMU), which substitute the transformer in the MulT model.

We have acquired two sequence representations for one modality which are given by $Z^{[D]}_{M_1 \to M_2 \rightrightarrows M_3}$ and $Z^{[D]}_{M_2 \to M_1 \rightrightarrows M_3}$. Following Tsai et al. (2019) and Rodríguez-Bribiesca et al. (2021), last time step (last token) is used for prediction. We are using a BERT encoder for modality L. BERT has a special token at the beginning called [CLS], which has all the information for classifying tasks. Thus, when we perform a CT projecting to the L space, we use the first token (time step) for classification because this is the corresponding attention of the [CLS] token.

Let be 'l

$$L_1 = L_{M_1 \to M_2 \rightrightarrows M_3} = \left[Z^{[D]}_{M_1 \to M_2 \rightrightarrows M_3}\right]_{S_{M_3}-1} \in \mathbb{R}^{1\times d}, \text{ and}$$

$$L_2 = L_{M_2 \to M_1 \rightrightarrows M_3} = \left[Z^{[D]}_{M_2 \to M_1 \rightrightarrows M_3}\right]_{S_{M_3}-1} \in \mathbb{R}^{1\times d}.$$

A concatenation similar to $[Z^{[D]}_{M_1 \to M_3}, Z^{[D]}_{M_2 \to M_3}]$ is used for prediction in both cases of Tsai et al. (2019) and Rodríguez-Bribiesca et al. (2021). Notice that they have one crossmodal projection level to modality M_3, so they need to find common patterns on each CT sequence output because these CT outputs come from different modalities (M_1 and M_2).

It is not the case with our architecture because L_1 and L_2 come from a similar set of low-level crossmodal projections, (M_1 adapted to M_2 and M_2 adapted to M_1).

Hence, instead of finding patterns between these CT outputs, we are fusing dynamically both L_1 and L_2 with a proposed fusion GMU module based on Arevalo et al. (2017). The proposed Fusion GMU (FGMU) is visualized in Figure 1. This module is described as follows using that $x_v = L_1$ and $x_t = L_2$:

$$h_v = \tanh(W_v x_v), \tag{9}$$

$$h_t = \tanh(W_t x_t), \tag{10}$$

$$z = \sigma\left(W_z[x_v, x_t]\right), \tag{11}$$

$$H_{M_3} = h = z h_v x_v + (1-z) h_t x_t, \tag{12}$$

where $W_v, W_t \in \mathbb{R}^{d\times d}, W_z \in \mathbb{R}^{d\times 2d}$ are the weights to be learned, and [,] is the concatenation operation.

In summary, we acquire a fused feature vector H_{M_3} for modality M_3, which is used to predict. To clarify, we are getting three different H_{M_3} vectors for each modality (L, V, A). These feature vectors are denoted as H_L, H_V, and H_A.

4.5. Simple Parallel Architecture

We have already explained the biprojection part of our proposed architecture **BPMulT**. In order to decrease the architecture depth effect, we also proposed a **simple parallel modality fusion** method that has inspiration in early, late, and hybrid fusions, Sourav and Ouyang (2021). This parallel architecture is considered the part that helps to not over-fit in the **BPMulT**. .

Let be $\overline{X}_M \in \mathbb{R}^{S_M \times d}$ the feature sequence for modality M as we have in the first projection. To make our model simple, we are reducing the sequence dimension of every modality to a fixed $S_m = 32$. This reduction is made by:

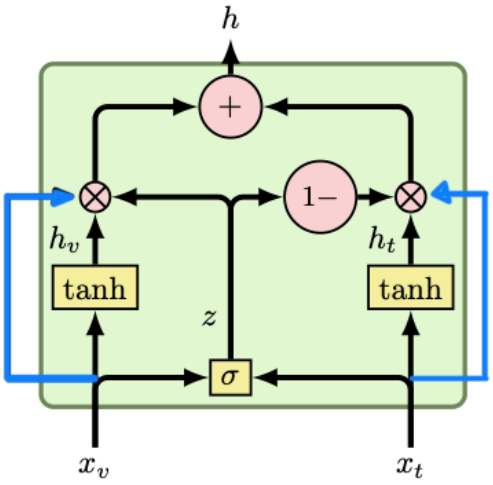

Figure 1: Proposed Fusion GMU (FGMU) for a dynamic weighted combination of two modality vectors x_v and x_t based on Arevalo et al. (2017). Blue lines are the proposed residual connections.

$$\widehat{X}_M = W_m \overline{X}_M, \tag{13}$$

where $\widehat{X}_M \in \mathbb{R}^{S_m \times d}$, and $W_m \in \mathbb{R}^{S_m \times S_M}$ a learnable weight matrix.

Then, we proceed to compute its positional embedding and we obtain a feature sequence $\hat{X}^m_M \in \mathbb{R}^{S_m \times d}$ defined analogously as in the first crossmodal section.

4.6. Self-Attention

For every $\hat{X}^m_M$ of each $M \in \{L, V, A\}$ we perform a self-attention based on Vaswani, Jones, Shazeer, Parmar, Uszkoreit, Gomez, Kaiser and Polosukhin (2017). This self-attention could be seen as the crossmodal attention having keys (K_M), queries (Q_M), and values (V_M) of a single modality.

Hence, we have the feature sequence attention defined as:

$$Y^m_M = \underset{M}{\mathbf{SA}}\left(\hat{X}^m_M\right) \tag{14}$$

$$= \text{softmax}\left(\frac{Q_M K_M^T}{\sqrt{d}}\right) V_M \tag{15}$$

$$= \text{softmax}\left(\frac{\hat{X}^m_M W_{Q_M} W_{K_M}^T \hat{X}^{m\,T}_M}{\sqrt{d}}\right) \hat{X}^m_M W_{V_M}, \tag{16}$$

where $Y^m_M \in \mathbb{R}^{S_m \times d}$, and $W_{K_M}, W_{Q_M}, W_{V_M} \in \mathbb{R}^{d\times d}$ are the weight to be learned.

Finally, we use $D = 3$ stacked self-attention layers to complete the transformer block as follows:

$$Z^{[0]}_M = \hat{X}^m_M, \tag{17}$$

$$\hat{Z}^{[i]}_{M_1 \to M_2} = \underset{M}{\mathbf{SA}}\left[\mathbf{LN}\left(Z^{[i-1]}_M\right), \mathbf{LN}\left(Z^{[0]}_M\right)\right] + \mathbf{LN}\left(Z^{[i-1]}_M\right), \tag{18}$$

$$Z^{[i]}_M = f_{\theta^{[i]}_M}\left[\mathbf{LN}\left(\hat{Z}^{[i]}_M\right)\right] + \mathbf{LN}\left(\hat{Z}^{[i]}_{M2}\right), \tag{19}$$

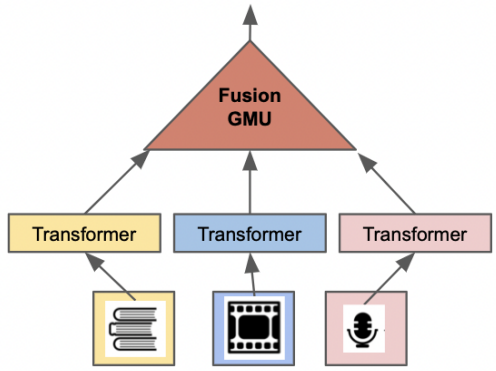

Figure 2: Proposed Parallel architecture with simple modalities representation and fusion with an FGMU extended to three modalities.

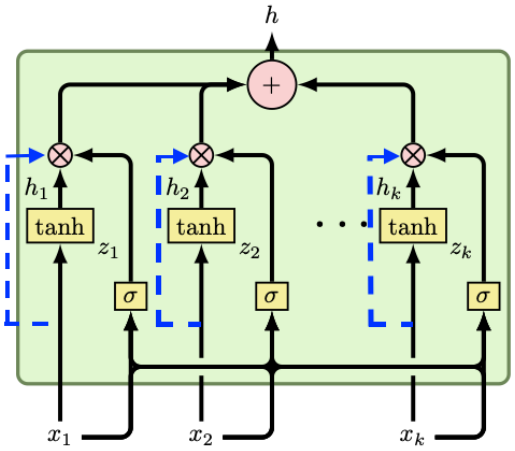

Figure 3: Proposed Fusion GMU (FGMU) extension for a dynamic weighted combination of three or more modality vectors $x_1, ..., x_k$ based on Arevalo et al. (2017). Blue lines are the proposed residual connections.

where $f_{\theta_M^{[i]}}$ and $LN()$ are defined the same way as before.

We now proceed to take the first sequence token of $Z_L^{[D]}$, L_l, and the last of $Z_V^{[D]}$, $Z_A^{[D]}$, L_v and L_a respectively, for classification. Note that every token is a d-dimensional vector.

4.7. Modalities Fusion with Extended FGMU

Since we have three d-dimensional representations (one of each modality), for simplicity of this parallel architecture, we are fusing each representation with the proposed **FGMU** in Figure 1 but extended for more than two modalities. The equations for this FGMU version, taking x_i as one of L_l, L_v, or L_a, are:

$$h_i = \tanh(W_i x_i), \tag{20}$$

$$z_i = \sigma\left(W_z^i[x_1, x_2, ..., x_k]\right), \tag{21}$$

$$H_P = h = \sum_{i=1}^{k} z_i h_i x_i, \tag{22}$$

where $W_i \in \mathbb{R}^{d\times d}, W_z^i \in \mathbb{R}^{d\times kd}$ are the weights to be learned for $i = 1, 2, ..., k$ (in this case $k = 3$), and [,] is the concatenation operation. With this method, we have a dynamically combined feature vector $H_P \in \mathbb{R}^{1\times d}$ representing the **simple parallel attention** of the multimodal sequences that are also used in our classification.

In Figure 2 we have an illustration of this simple method proposed, and the proposed **FGMU** extension can be viewed in Figure 3.

4.8. Dynamic Modalities Fusion with GMU

In Section 4.4, we have obtained the H_L, H_V, and H_A feature vectors representing a modality summary that takes into account the attention given between the other modalities. In Section 4.5, a simple summary vector H_P was computed to recover general and relevant information. Furthermore, depending on the taken dataset, we might have extra non-sequential important features we want to use to classify, e.g., images or metadata. To address this wanted fusion, we worked with a GMU module as it is proposed in Arevalo et al. (2017).

We mainly take our modality feature vectors (H_L, H_V, H_A) and our simple summary vector H_P into the GMU. Extra relevant features are denoted as $H_E \in \mathbb{R}^{1\times d}$. These features are dynamically combined with the original GMU explained in Section 2. The GMU module gives an output $O \in \mathbb{R}^c$, where c is the number of classification categories.

Hence, this process is described as:

$$\begin{aligned}
h_1 &= \tanh(W_1 H_L),\\
h_2 &= \tanh(W_2 H_V),\\
h_3 &= \tanh(W_3 H_A),\\
h_4 &= \tanh(W_4 H_P),\\
h_5 &= \tanh(W_5 H_E),\\
z_i &= \sigma\left(W_z^i[H_L, H_V, H_A, H_P, H_E]\right), \text{ for } i = 1, 2, 3, 4, 5,\\
O &= h = \sum_{i=1}^{5} z_i h_i,
\end{aligned} \tag{23}$$

where $W_{1,2,3,4,5} \in \mathbb{R}^{c\times d}$ and $W_z^i \in \mathbb{R}^{c\times 5d}$ are learnable weights.

After dynamic fusion, we perform a traditional residual block before giving the final prediction:

$$O_f = W_{f_1}\left[\text{dropout}(\text{ReLu}(W_{f_2} O))\right] + O, \tag{24}$$

where $W_{f_1}, W_{f_2} \in \mathbb{R}^{c\times c}$ are weights to be learned.

4.9. Tackling Vanishing Gradient

This section is also one of the essential parts of the proposal because it involves the solution to our third specific motivation.

One of the most common problems in deep neural networks is the vanishing gradient. Our proposed architecture has three depth primary levels where residual connections are intuitive between levels and is a potential tool for preventing this inconvenience.

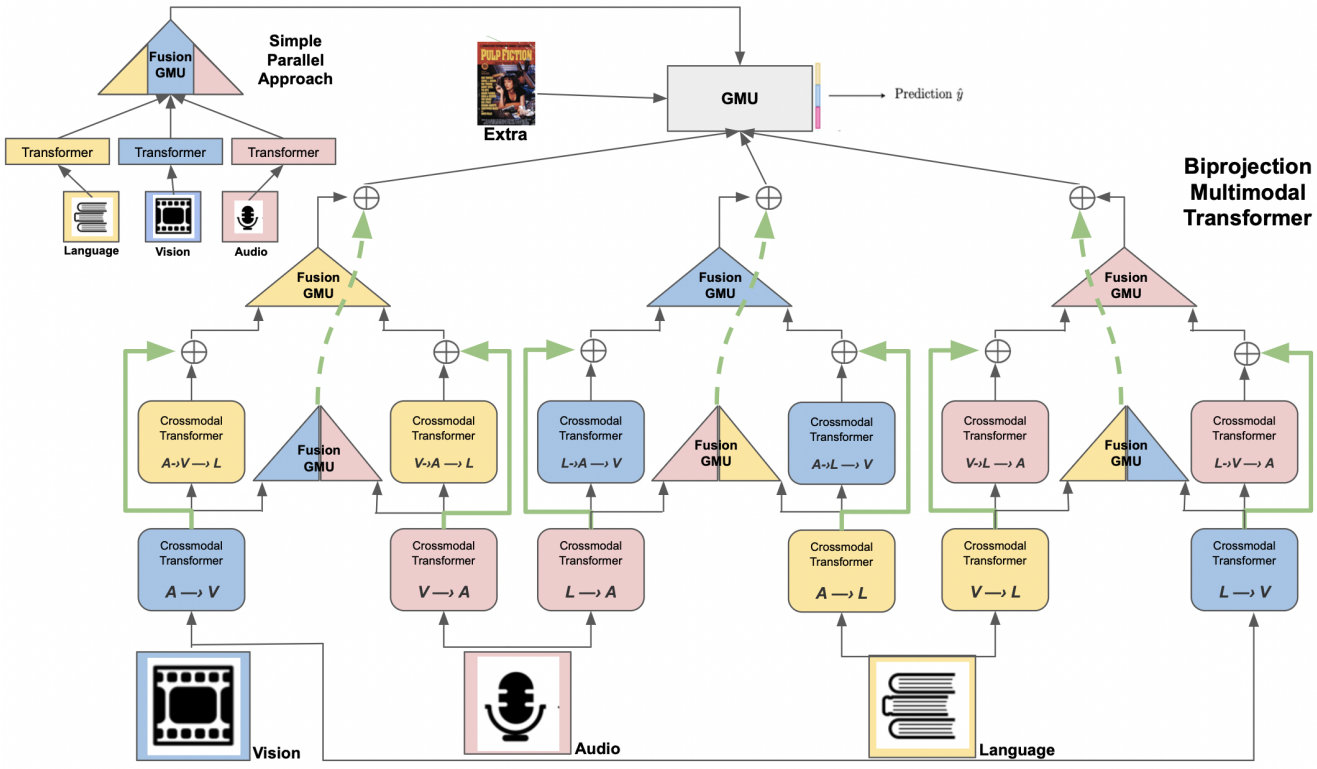

Figure 4: Proposed Biprojection Multimodal Transformer **BPMulT** architecture for multimodal classification tasks. Yellow blocks correspond to text modality projected Crossmodal Transformers (CT), blue blocks correspond to Video modality projected CT, and red blocks correspond to audio modality projected CT. Its color represents the FGMU as the modalities that they are fusing. The GMU in block color gray weighs to combine the information from a simple parallel architecture, a heavy model, and extra metadata features. The simple parallel part is introduced to reduce the over-fitting of the heavy architecture.

We introduce two kinds of residual connections to the model. One is a connection between the first crossmodal projections 4.1 and the second 4.2 just at a single token level (specific token used for prediction). Following the preceding notation:

$$L'_1 = \left[Z^{[D]}_{M_1 \to M_2 \rightrightarrows M_3}\right]_{S_{M_3}-1} + \left[Z^{[D]}_{M_1 \to M_2}\right]_{S_{M_2}-1} \in \mathbb{R}^{1\times d}, \text{ and} \quad (25)$$

$$L'_2 = \left[Z^{[D]}_{M_2 \to M_1 \rightrightarrows M_3}\right]_{S_{M_3}-1} + \left[Z^{[D]}_{M_2 \to M_1}\right]_{S_{M_1}-1} \in \mathbb{R}^{1\times d}. \quad (26)$$

The second residual connection is a token prediction fusion with an FGMU of the first crossmodal projection to the prediction token fusion of the second crossmodal projection. Respective equations are:

$$H'_{M_3} = \textbf{FGMU}\left[Z^{[D]}_{M_1 \to M_2}\right]_{S_{M_2}-1} + H_{M_3} \in \mathbb{R}^{1\times d}. \quad (27)$$

An illustration of our proposed **BPMulT** model for multimodal classification can be observed in Figure 4 and the proposed compounding parts in Figure 7.

5. Experimental Framework

Following previous work (MulT-GMU, Bribiesca 2021), we compare our two base models with the MulT-GMU on Moviescope, which was also compared with the Multimodal BERT (MMBT) and Fast Modal Attention (Fast-MA). For the MM-IMDb dataset, we include the actual results presented in (Arevalo., 2017), and for IEMOCAP and MOSEI, we have the original SOTA results.

We denote different modalities as V (Video), A (Audio), P (Poster), and T (Text). In the case of the MM-IMDb dataset, we handle V, A, and T as three different text embeddings since this dataset only has text and poster modalities. For our models BPMulT and BPMulT-no-parallel, we show their mean and standard deviation of the performance over five runs with random seeds for all datasets considered.

We select the best hyperparameters for every dataset, including the number of heads, hidden dimension, number of layers in the transformer, batch size, and gradient accumulation step.

We use the Moviescope dataset **??** to determine the best model configuration (in terms of architecture) for all the experiments below. With this final model, we perform the experiments in the other datasets.

6. Experiments in Moviescope

The way modalities interact in the multimodal transformer model is the base for developing our architecture. The motivation to create a novel transformer-based architecture is to enrich each modality with information from the others.

Modalities	Model	μAP	mAP	sAP
TVAPM	MulT-GMU	79.5±0.5	76.4±0.3	85.6±0.3
	MulT-GMU-no-transf-encoder	80.3±0.2	76.9±0.2	86.1±0.4
TVAP	MMBT	77.4±0.7	74±0.8	85.1±0.7
	Fast-MA	74.9	67.5	82.3
	MulT	78.9±0.3	75.7±0.5	85.6±0.3
	MulT-GMU	79.8±0.4	76.0±0.9	86.1±0.4
	BPMulT-no-parallel (ours)	**81.4±0.3**	**78.0±0.5**	**87.2±0.4**
	BPMulT (ours)	**81.7±0.2**	**78.1±0.1**	**92.4±0.1**

Table 1
Comparison of our BPMulT model with different modality combinations. Modalities refers to text-video-audio-poster (TVAP) and text-video-audio-poster-metadata (TVAPM). The models use just the specified modalities' information. Metrics reported corresponding to average precision, micro (μAP), macro (mAP) and sample (sAP) averaged.

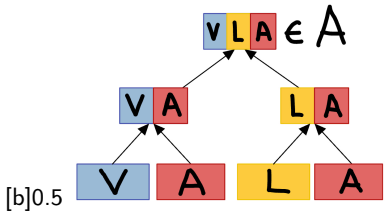

Figure 5: Single modalities combinations to get an enriched modality A.

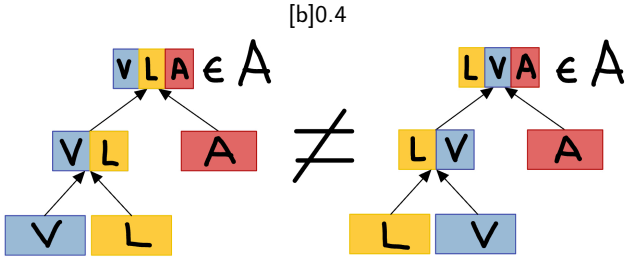

Figure 6: Joint modalities combinations to get an enriched modality A.

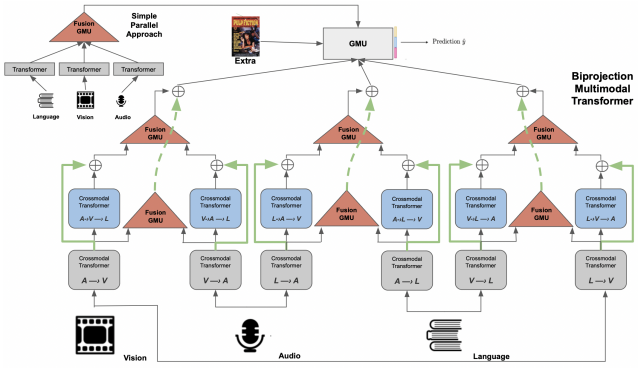

Figure 7: Proposed **BPMulT** architecture for a multimodal classification task. The blue part corresponds to biprojection modules. Red triangles are the GMU modules for fusion. Green arrows correspond to residual connections from the first projections. Gray modules are the first crossmodal transformers. We can observe the parallel reduced fusion step in the upper left corner. The GMU weighs the heavy and simple architectures and the metadata features to give a classification.

The primary purpose of this experiment is to improve the best results shown in (Rodriguez, 2021) and its baselines.

We have the intuition that the **joint projection between two modalities with a third modality is better than just a single projection of one modality to another** as we can see in Figure 6 the joint combination and in Figure 5 the single combination.

We first compare the Moviescope baseline models, Fast-MA, MulT, and MulT-GMU, with our BPMulT with and without parallel fusion in the final GMU. Results on the modality setting (TVAP) are shown in Table 1. Note that our proposed models outperform the SOTA models by 2% on each metric even when the modality setting of the MulT-GMU is TVAPM (with metadata).

In Figure 7, we observe our proposed BPMulT model with the reduced parallel fusion to not overfit the dataset. The BPMulT-no-parallel model is the same architecture removing the upper left corner part (the reduced parallel fusion).

6.1. BPMulT Ablation Experiments

We propose a set of experiments to evaluate the main components of our BPMulT model where we remove each one of them. These main components correspond to the FGMU modules for fusion information of a single modality and the residual connections proposed through the crossmodal transformers' first and second levels. We show only the main results of the ablation study to measure the impact on the model's performance.

No-FGMU modules. We propose fusing one modality's different information with an FGMU, e.g., fusing FGMU(A—›L and V—›L). To obtain the relevance of this FGMU module, we follow the idea of (Rodríguez, 2021) removing it and passing the information directly to the GMU.

No-FGMU modules nor Residual Connections. In order to obtain the relevance of the residual connection proposed, we removed them with this experiment.

All Crossmodal Transformers to GMU. In this experiment, we are looking for details about if a biprojection is useful for the classification. Following (Rodríguez, 2021),

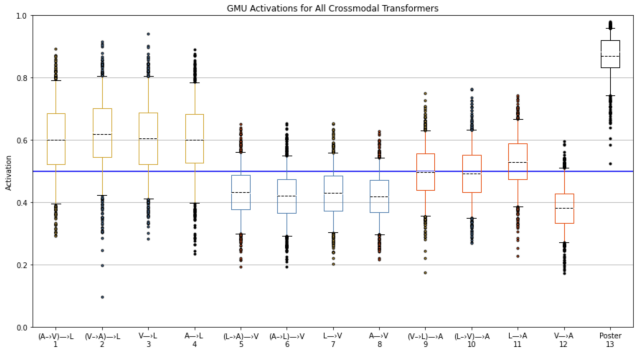

Figure 8: GMU activations of the experimental proposed architecture passing all crossmodal transformer encoders (biprojections and single projections) for ablation study. Yellow bars correspond to activations of the Language space, blue ones to the Video space, red ones for audio and black ones to the poster features.

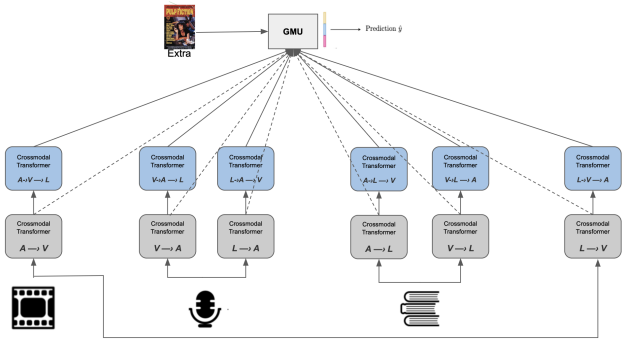

Figure 9: Proposed **BPMulT-no-parallel** architecture for a multimodal classification task passing all transformer encoders (biprojections and single projections) for ablation study. The blue part corresponds to biprojection modules. Gray modules are the first crossmodal transformers.

the best performance of the MulT-GMU model is when all crossmodal (single projection in gray) are given to the GMU. So, we are giving all the information from the biprojections and the single projections to the top GMU, as shown in Figure 9.

In Figure 8, we can note that the activations of the biprojections at the GMU are similar to or better than a single projection. For example, in the Text space, projecting from Video (third bar) is almost the same as the transition Audio to Video and then to text (first bar). On the other hand, the projection of Video to Audio and then to text (second bar) is better than just projection from audio (fourth bar).

No-Residual Connections. This experiment investigates if the residual links from the first crossmodal projections to the biprojection are helpful for the model.

No-FGMU in the Middle. Similaly, we investigate if the Fusion GMU in the middle is relevant for the model. The Middle FGMU fuses the first crossmodal projections to add them to the last biprojection.

Transformer Attention (concat). We designed the following ablation experiment to compare if the fusion with a self-attention transformer is better than our proposed fusion. The architecture is substituting the FGMU at the top with a transformer that receives a concatenated vector.

Cross-Attention Modification (Translating)

For this experiment, we investigate the crossmodal attention of the biprojection at its nucleus. This crossmodal transformer was proposed in Tsai et al. (2019), and we use it with other modalities data in a different order to obtain better sequence patterns. In Figure 10, we can observe how we are doing the biprojection step replicating the crossmodal transformer for the second attention.

Our intuition was to follow the idea of the original transformer, which was first used as a Neural Machine Translation. We believe that we can handle modalities as languages. In other words, we want to translate the sequence in "language" called **text** to another language called **Video**.

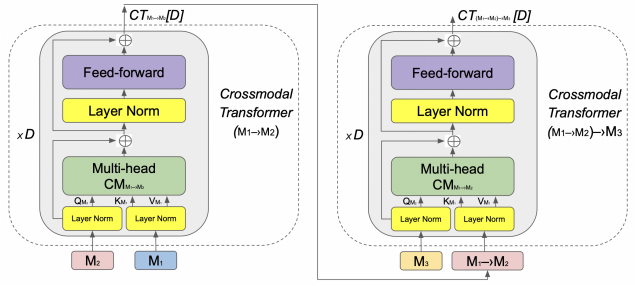

Figure 10: Proposed **Biprojection** mechanism composed by two crossmodal transformers. It can be viewed as an encoder-decoder between modalities.

Following Vaswani et al. (2017), we propose (for ablation experiment) an attention inspired in a translation. The modification only replicates the "decoder" part of the transformer in our biprojection. It is just adding a self-attention module followed by a normalization layer. We can observe the biprojection as a decoder module in Figure 11.

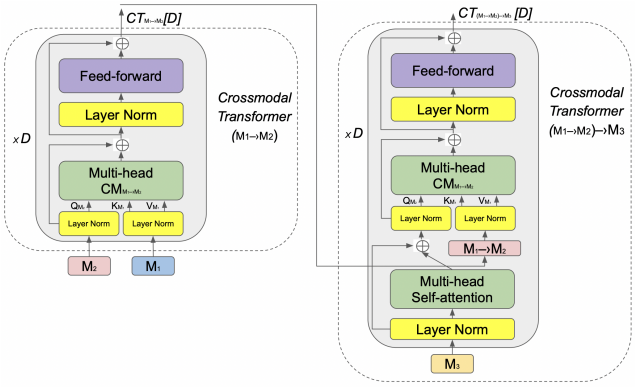

Figure 11: Proposed **Biprojection** mechanism with a modification for translating. It comprises a crossmodal transformer as an encoder and another crossmodal transformer with a self-attention and a normalization layer as a decoder.

Modalities	Model	μAP	mAP	sAP
TVAP	BPMulT-translating (ours)	81.2±0.3	77.8±0.5	87.0±0.3
	BPMulT-with-transformer (ours)	79.97±0.1	76.6±0.3	86.2±0.3
	BPMulT-no-FGMU-Middle (ours)	80.95±0.3	76.6±0.5	86.9±0.4
	BPMulT-no-RC (ours)	81.2±0.3	77.7±0.5	87.1±0.3
	BPMulT-no-FGMU (ours)	79.9±0.25	76.54±0.3	86.2±0.4
	BPMulT-no-FGMU-nor-RC (ours)	80.5±0.5	76.9±0.5	86.4±0.3
	BPMulT-all-transformers (ours)	79.8±0.4	76.4±0.3	86.1±0.5
	BPMulT-no-parallel (ours)	**81.4±0.3**	**78.0±0.5**	**87.2±0.4**
	BPMulT (ours)	**81.7±0.2**	**78.1±0.1**	**92.4±0.1**

Table 2
Comparison of our BPMulT model with all outperformed ablation experiments. Metrics reported corresponding to average precision, micro (μAP), macro (mAP), and sample (sAP) averaged.

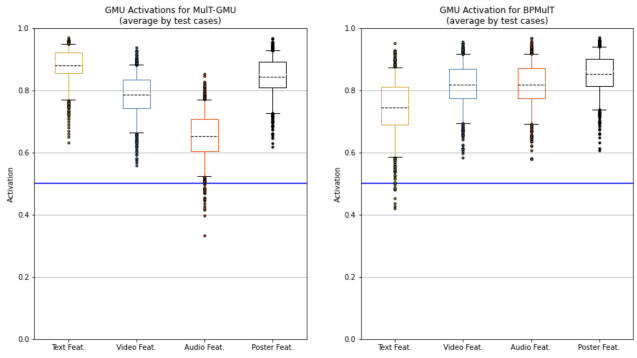

Figure 12: Comparison of features activation by the GMU for dynamic fusion for the Moviescope dataset (Cascante-Bonilla et al., 2019). On (left), we can observe the activation given by the model MulT-GMU of Rodríguez-Bribiesca et al. (2021), and on the (right), we show the activation of our proposed model BPMulT.

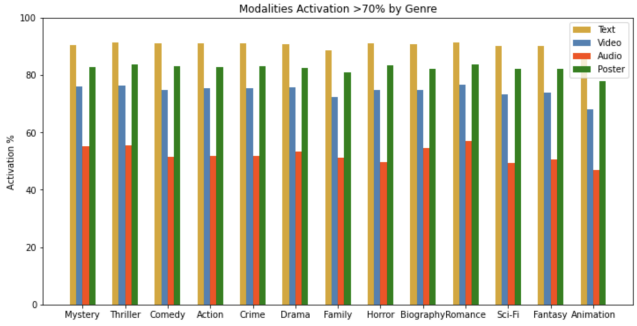

Figure 13: Label activation of the GMU for dynamic fusion for the Moviescope dataset (Cascante-Bonilla et al. (2019)) given by the model MulT-GMU of Rodríguez-Bribiesca et al. (2021) by genre.

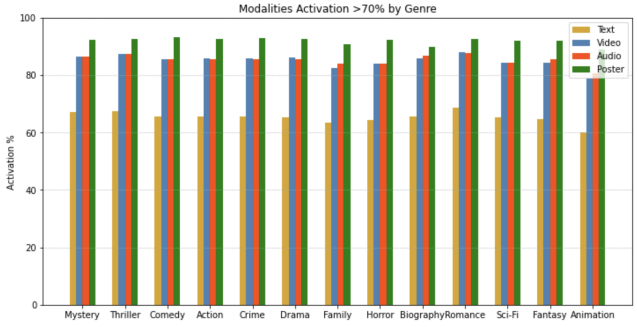

Figure 14: Label activation of the GMU for dynamic fusion for the Moviescope dataset (Cascante-Bonilla et al. (2019)) given by our proposed model BPMulT by genre.

7. Relevance of Modalities

This section aims to understand which modalities the model considers for predicting. We are analyzing the GMU module activations on the test set and the FGMU activations to see the dynamic flow of information in the model. In addition, we complement the analysis with a TSNE study. This TSNE study shows the vectors' data colored by their genre and plots them in a two-dimensional space by its predicted label in the test case. With this visualization, we can see the data grouped by genre (same color) if the model is doing a correct classification.

7.1. GMU Activations

To compare our model, we first analyze the activation that the MulT-GMU of Rodríguez-Bribiesca et al. (2021) has. On the left of Figure 12 we can observe the modalities feature activation of the MulT-GMU, the most activated modality is text followed by the poster. It shows that text has more relevance for classification than audio or video modalities. The video modality is, on average activated 5% less than the Text modality, and the audio is 10% less activated than text.

Note that the MulT-GMU has various activations for each modality. Remarkably, the text modality is notably less activated with our proposed model. Poster features are still activated for more than 80%. Video and audio modalities are now activated with a big difference. This section aims to find why this is happening and whether this modality is enriched with better sequence patterns or whether they have information that the BPMulT uses, and the MulT-GMU does not.

In Figure 13, we can observe the GMU activations for each modality but separated by genres. We note that these activations are not the same for all genres, but they are similar.

The same case we have for the BPMulT activations is in Figure 14. The average activation is similar for all the genres, and the less activated modality is text. Video and audio are almost equal activated, and the poster is more used for the classification.

7.2. TSNE Study of GMU Activations

To complement this analysis, we do a TSNE study of the GMU's activation for each modality in the BPMulT model. If one modality is more relevant for the model than the others, it is more comprehensive and may have a better clusterization with this study. In Figure 15 we can observe the activation of the text modality group labels.

Furthermore, there are clusters in the space. We have the Comedy (blue) and Drama (red) data on one side. Between them, we have some Romance (dark green) movies. On the other side, we can observe ocher and yellow clusters

corresponding to Fantasy and Animation genres. Next to them is a pink cluster of Science-Fiction movies, and in the green set, we have Horror movies close to the black cluster corresponding to the Crime genre. Note that these mentioned movies are related, and the clusters make sense for us as human classifiers. It is a good signal that our model is doing well for classification.

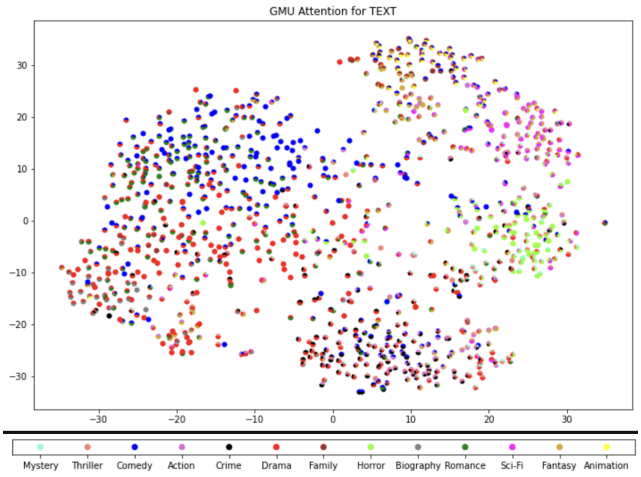

Figure 15: TSNE study of the output of the last GMU for a dynamic fusion of modalities in the BPMulT model. Each color corresponds to a genre, and data is multi-labeled. We are visualizing the activations corresponding to the **text** part.

Similarly, the activations of video, audio and poster modalities in Figures **??**, **??**, **??** in the Appendix **??** are well clustered being the audio and video the modalities with the best performance in activation and with clear clusters. If there is a perfect clusterization of the data, the model performs better since it knows how to classify perfectly by genre.

7.3. Understanding the FGMU Activation Flow

Now, we are interested in analyzing the flow of information in the FGMUs to get the actual contribution of each modality. In Figure 16, we compare the GMU activations when we fuse the first projections and the biprojections. The first row corresponds to all single projections from one modality to another. The second row has the activations of the biprojections. We can understand the information's flow in the model considering the graphs ordered by columns.

In the first box plot, we have that the features of a projection from Audio to Video are more activated than Video to Audio. Then, below the first graph, we have that the biprojection features of (A–›V)—›L are used more than the features of (V–›A)—›L, which has accordance with the relevance in the first graph. Hence, features of (A–›V) and (A–›V)—›L are mainly used to represent the Text modality. Figure 12 shows that the Text modality is not as well activated as the Audio and Video modality in the final GMU. It could be a consequence of using a large amount of Video information.

For the Video representation, note that projections and biprojections are equally used, which could be the reason for a high activation of this modality. We will discuss this case with other studies to confirm this idea.

Finally, the Audio representation is compounded by mostly Text and (V–›L)—›A features. The Audio representation also has a high activation. As a first conclusion, the features with more relevance to the classifications are from (V–›L)—›A and V–›L coming from the Audio modality. For the Video modality, the information comes from A—›L, (A–›L)—›V, L—›A, and (L–›A)—›V.

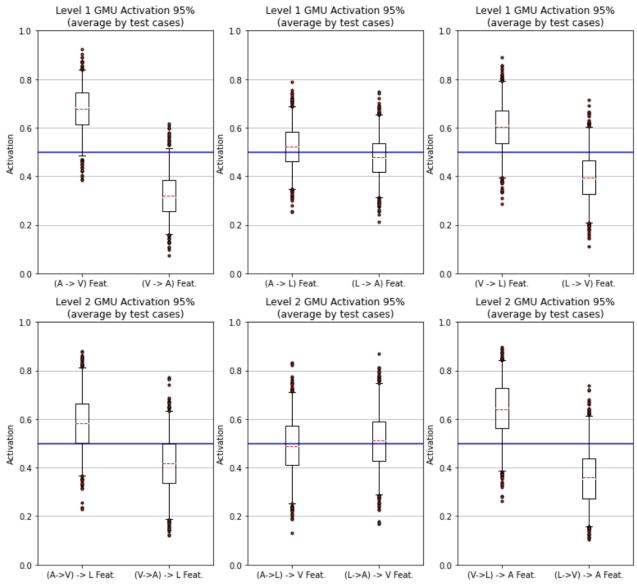

Figure 16: Comparison of features activation by each FGMU (Top and Middle) for dynamic fusion for the Moviescope dataset (Cascante-Bonilla et al., 2019). In the **first** column we can observe the activation given by the BPMulT to get the **text** modality representation, in the **second** column the activation to get the **video** modality representation, and in the **third** column the activation of the **audio** representation. The first row corresponds to the Middle FGMU and the second row to the Top GMU.

7.4. TSNE Study of the Dynamic FGMU Activation Flow

In order to investigate why the FGMU modules are taking the information flow as we have seen in Figure 16, we analyze the Crossmodal Transformers involved with a TSNE study. We show below an example analysis of the text branch modality representation. The text branch refers to the BPMulT part used to represent the text modality, and we can see this branch in Figure 17. All the other TSNE study graphics for the other modality branches are in Appendix **??**.

In Figure 18, we see that the data distribution is not as accurate as we desired. We can not identify any group with this TSNE graph. In contrast, Figure 19 shows a great change in data distribution. We can easily identify the cluster of Horror, Sci-Fi, Drama, or Comedy movies. Hence, this

biprojection should be strongly used to represent the text modality.

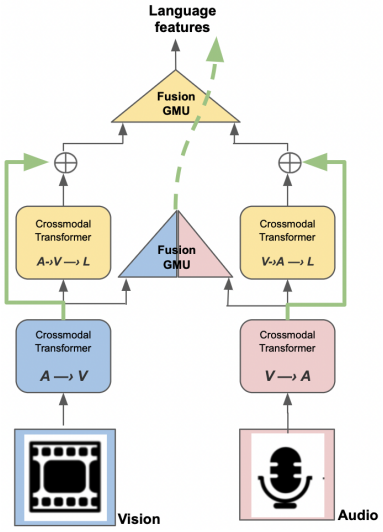

Figure 17: Text branch of the Biprojection Multimodal Transformer (BPMult) model. It consists of two opposite projections ((A—›V) and (V—›A)) and their respective biprojections to the text space.

On the other branch of the Text representation, we have the projections V—›A and (V–›A)—›L. The corresponding TSNE study of each projection is in Figure 20 and 21. We can observe in these figures something similar to the other branch, i.e., the projection V—›A is not clearly clustered but biprojecting this to the Text space makes the distribution better to classify.

Remembering that in our proposed model, the Text modality is not well activated (Figure 12), and to represent this modality is used with a majority of the A—›V and (V–›A)—›L features (Figure 16). We conclude that the activation of the Text modality is disadvantaged by using a large amount of A—›V features.

Doing an analogous analysis for Audio and Video modalities, we found that V—›L and (V–›L)—›A are well clustered, and it is because the Audio modality has a high activation. In the case of Video, the A—›L and L—›A are well grouped (the biprojections to the Video modality not). Still, we have the same activation for the four modules, so the modality of Video is also well activated.

With this analysis, we obtained the exact flow of information. It showed us why modalities like Video and Audio are enriched and the text modality's low activation. However, with the residual connections between levels, all types of crossmodal transformers are considered for the final prediction. They help to find better patterns and to improve the overall score.

7.5. BPMulT in the MM-IMDb Dataset

We have mentioned that MM-IMDB has just two modalities for information extraction, text (synopsis) and its poster. Since BPMulT works finding patterns between sequences.

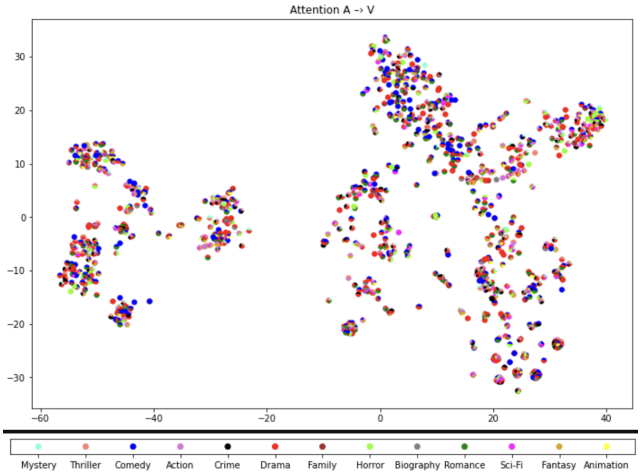

Figure 18: TSNE study of the output of the crossmodal transformer (A—›V) in the BPMulT model. Each color corresponds to a genre, and data is multi-labeled. It is considered a **bad** clusterization.

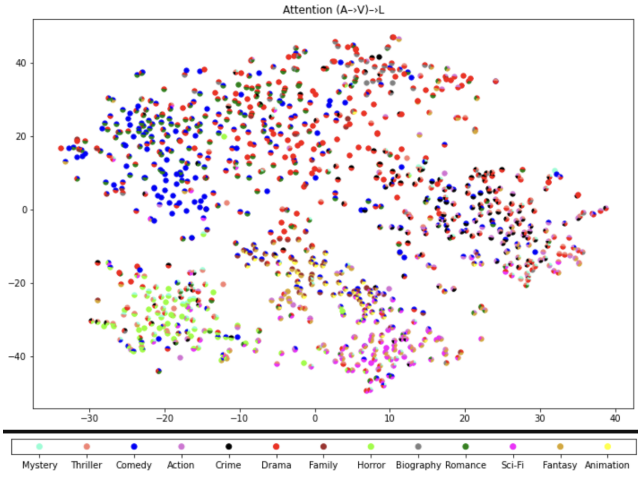

Figure 19: TSNE study of the output of the crossmodal transformer with biprojection (A—›V)—›L in the BPMulT model. Each color corresponds to a genre, and data is multi-labeled. It is considered a **good** clusterization.

We proposed to represent the synopsis with various text embedding. Our first approach is to use the BERT representation because we use the transformer architecture. The other representations selected are the GloVe embedding which is a widely used text embedding, and the third one is just simple BoW taking as vocabulary the ten thousand most frequent words used in the training set. The poster will be fused with a GMU at the last part of the BPMulT model since it is not a piece of sequential information.

Table 3 shows the results for the metrics considered in past publications. The first model proposed is the GMU which was overpassed before by the MMBT model. We performed five random training seeds to get the mean and standard deviation metrics for the BPMulT model. We achieved better metrics in μ-F1, W-F1, and s-F1 than reported for the previous SOTA models.

MM-IMDB				
Model	μ-F1	m-F1	W-F1	s-F1
GMU	63.0	51.4	61.7	63
MulT-GMU	66.3±.6	61.1±.6	-	-
ConcatBERT	65.9±.2	60.5±.3	-	-
MMBT	66.8±.1	**61.6±.2**	-	-
BPMulT-no-parallel (ours)	**68.9±.1**	58.7±.3	**68.8±.1**	**69.4±.2**
BPMulT (ours)	**68.9±.1**	58.7±.3	**68.8±.1**	**69.4±.2**

Table 3
Comparison of our BPMulT model in the MM-IMDb dataset where the model MMBT is the current SOTA model. Metris reported are F1 scores micro (μ), macro (m), weighted (W), and sample (s) averaged.

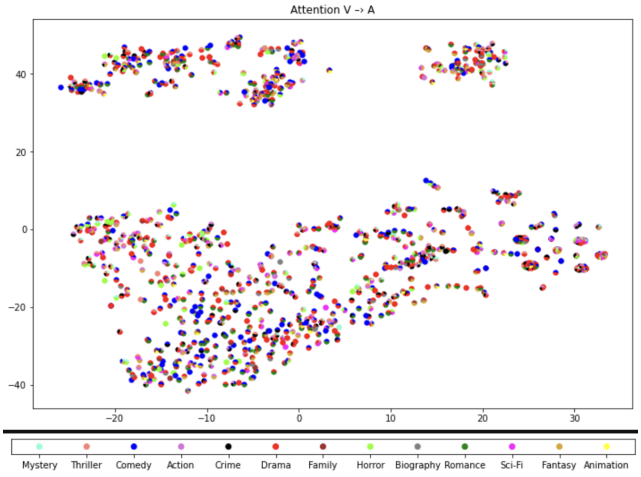

Figure 20: TSNE study of the output of the crossmodal transformer (V—›A) in the BPMulT model. Each color corresponds to a genre, and data is multi-labeled. It is considered a **bad** clusterization.

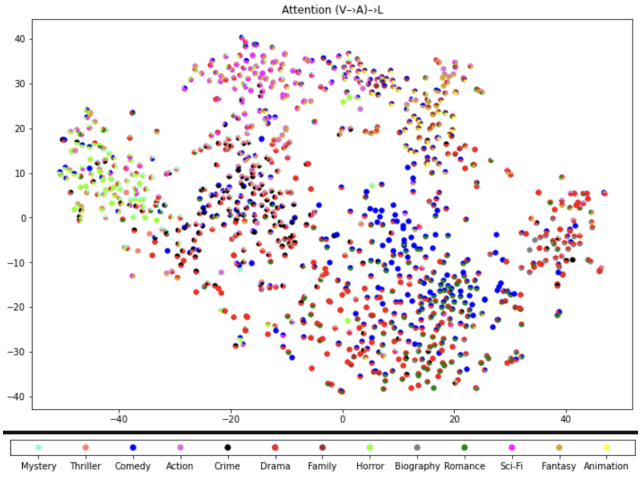

Figure 21: TSNE study of the output of the crossmodal transformer with biprojection (V—›A)—›L in the BPMulT model. Each color corresponds to a genre, and data is multi-labeled. It is considered a **good** clusterization.

7.6. BPMulT in the IEMOCAP Dataset

IEMOCAP is one of the most used datasets to test multimodal architectures since it has three modalities for information extraction. It has features from Text (GloVe), Audio, and Video. With this dataset, BPMulT works directly to find patterns between sequences. We do not have any non-sequential modality here. Then, we fuse all three modalities with a GMU at the last part of the BPMulT. IEMOCAP provides an aligned sequence of length **twenty** since the text modality is sectioned in twenty parts. We consider this alignment in Table 4 for our first approach, but it is unnecessary. In the Table 5 we can see the classifications result for unaligned sequences.

Table 3 shows the results for the metrics considered in Dai, Liu, Yu and Fung (2020). We took this publication as the SOTA model because it considers that the F1 metric is inappropriate for unbalanced datasets. To avoid this problem, they propose using the AUC score. We performed random seed training for the BPMulT model. We achieved a better AUC metric and a competitive score in the Accuracy metric.

We performed random seed training for the BPMulT model. We achieved a better Accuracy score for Happy, Sad, and Angry emotions, which is an excellent result because the class with more labels is the Neutral emotion. On the other hand, Table 3 shows the results for the metrics considered in Tsai et al. (2019) because it is the first model used for unaligned sequences.

7.7. BPMulT in the CMU-MOSEI Dataset

CMU-MOSEI is the most helpful dataset to test multimodal models for emotion and sentiment analysis. It also has three modalities for information extraction. Its features came from Text, Audio, and Video. Similar to the IEMOCAP dataset, BPMulT works directly, and we fuse all three modalities with a GMU in the last part. For this version of the CMU-MOSEI, we took the proposed modification of Dai et al. (2021), which provides a reordering and cleaning of some conversations. Hence, the SOTA model corresponds to the proposed models in the same publication (FE2E and MESM) for **unaligned** sequences. Since this dataset provides raw text and we are working with crossmodal transformers, we select BERT as our encoder for text representation. Audio and Video representation are the provided vectors by the mentioned publication.

Table 6 shows the results for the metrics considered in Dai et al. (2021). We performed five random training seeds for the BPMulT model to obtain the reported metrics' mean

IEMOCAP Aligned

Model	Neutral		Happy		Sad		Angry		Average	
	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC	Acc	AUC
MulT	71.0	77.2	83.5	71.2	85.0	89.3	85.5	92.4	81.3	82.5
ModTrans-MMEmoRe	71.1	76.7	85.0	**74.2**	**86.6**	88.4	**88.1**	**93.2**	**82.7**	83.1
BPMulT-no-p (ours)	**71.9**	**78.4**	**87.0**	73.0	85.8	**89.0**	85.5	92.3	**82.6**	**83.2**
BPMulT (ours)	65.0	70.3	85.5	73.5	78.8	76.4	81.1	85.6	77.6	76.4

Table 4
Comparison of our BPMulT model in the aligned IEMOCAP dataset where the model ModTrans-MMEmoRe is the current SOTA model corresponding to the model in Dai et al. (2020) for emotion recognition. Metrics reported are the accuracy and the Area Under the Curve (AUC) instead of F1 as Dai et al. (2020) proposed.

IEMOCAP Unaligned

Model	Neutral		Happy		Sad		Angry		Average	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
MulT	**62.5**	**59.7**	84.8	**81.9**	77.7	**74.1**	73.9	**70.2**	74.7	**71.5**
BPMulT-no-p (ours)	59.6	45.7	**85.6**	79.0	**79.4**	70.3	**75.8**	65.4	**75.1**	65.1
BPMulT (ours)	59.2	44.0	**85.6**	79.0	**79.4**	70.3	75.6	66.2	74.9	64.9

Table 5
Comparison of our BPMulT model in the not aligned IEMOCAP dataset where the model MulT is the current SOTA model since this model was proposed for unaligned sequences. The metrics reported are the accuracy and the F1 score.

and standard deviation scores. We achieved better Weighted-Accuracy in almost every emotion.

8. Conclusions

Accordingly to our objectives in Section **??**, we have four specific objectives to achieve and one main goal, which is the primary purpose of our research. Following the structure of our specific motivations, we these conclusions of our work:

1) Our first specific objective is to improve the modalities' combination. We proposed a novel architecture that involves a biprojection that enriches each modality representation with information from the other modalities. Ablation experiments and results show that this biprojection is crucial to rescue relevant information from the sequences since it also allows residual connections and intermediate fusion modules to be possible. Hence, we achieve our first goal with the proposed crossmodal biprojection.

2) The second specific objective is to substitute the heavy information fusion method (the transformer). We proposed the Fusion GMU (FGMU) module, a lower-cost method that does not decrease the performance in classification. Ablation experiments and results show that the FGMU helps the model to improve its classification performance. It also is easy to interpret because it learns to weigh each modality to fuse them. We can see the used weights and interpret them as feature activations. Then, we achieve our second goal with this proposed module.

3) The third specific objective is to improve learning by introducing strategically residual connections. Since we proposed to introduce the crossmodal biprojections to our model, we facilitate addressing this problem because we can use the information from the first projection and enrich each modality. Using the FGMU also helped to place better connections linking crossmodal blocks' summarized information. The ablation experiment shows that our proposed configuration of connections has the best performance. Hence, we satisfied our third goal with the connections of biprojections and FGMU modules.

4) Finally, our fourth specific objective is to understand the information flow in the proposed architecture. With the FGMU, we achieved this goal because it provides an interpretable weighing system. We obtained the following conclusions with the activation of the GMU and FGMU modules. With the TSNE analysis, we see the excellent clusterization of Moviescope test labels when we see its FGMU activation in the projections and biprojections. We detect that in the branch of the text features (Figure **??**), the projections are not well clustered, and the biprojections are. With the FGMU activation analysis in Section 7.3, we see that the information (Figure 16) is mainly used from (A—›V), which is poorly clustered and (A—›V—›L) which is good. It results in a low activation for the text modality in the final GMU prediction, accordingly to Figure 12.

CMU-MOSEI Unaligned

Model	Angry W-Acc/F1	Disgust W-Acc/F1	Fear W-Acc/F1	Happy W-Acc/F1	Sad W-Acc/F1	Surprised W-Acc/F1	Average W-Acc/F1
MulT	64.9/47.5	71.6/49.3	62.9/25.3	67.2/**75.4**	64.0/48.3	61.4/25.6	65.4/45.2
FE2E	67.0/**49.6**	**77.7**/57.1	63.8/26.8	65.4/72.6	65.2/**49.0**	**66.7**/**29.1**	67.6/**47.4**
MESM (0.5)	66.8/49.3	75.6/56.4	65.8/28.9	64.1/72.3	63.0/46.6	65.7/27.2	66.8/46.8
BPMulT-no-p (ours)	66.7/26.5	69.0/**75.6**	**66.2**/**48.7**	**74.7**/50.8	**67.3**/48.5	62.9/26.5	**68.3±.3**/46.1±.4
BPMulT (ours)	**69.3**/26.4	68.0/**74.9**	**66.0**/**48.5**	**74.5**/50.9	**67.1**/48.3	63.4/25.1	**68.0±.6**/45.7±.6

Table 6
Comparison of our BPMulT model in CMU-MOSEI unaligned dataset modified by Dai et al. (2021) where the model FE2E is the current SOTA model proposed in the same mentioned publication. The metrics reported are the weighted accuracy (W-Acc) and the F1 score.

In contrast, in the branch of the video features, the projections are well clustered, but the biprojections are not. The FGMU activation analysis shows that the information (Figure 16) is equally activated and is used even if it is good and bad clustered data. It results in a high, but not highest, activation for the video modality in the final GMU prediction, following Figure 12.

Moreover, in the branch of the audio features, something different happens: one projection is good activated as well as its biprojection, and the other is poorly clustered and also its respective biprojection. The information mainly corresponds to the good clustered data with the FGMU activation analysis (Figure 16). It results in the highest activation, the audio modality in the final GMU prediction, accordingly to Figure 12.

In conclusion, biprojections and single projections are relevant if we only see the clustering with the TSNE study. The advantage of the BPMulT model is that it dynamically takes into account features of both single crossmodal transformers and biprojections. However, an ablation experiment shows that the biprojections have more activation than projections, and in consequence, biprojection can be seen as more relevant.

Finally, our main objective is to improve the representation of modalities' combinations within the Multimodal Transformer architecture. Also, we aim to achieve superior performance to the SOTA models in movie genre classification and emotion recognition tasks. The BPMulT model improves the representation of each modality with information from the other modalities. It is done with the help of biprojections and fusion modules. Results show that the activation increase for modalities not relevant before. Our proposed BPMulT model has mainly achieved our goal and obtained the SOTA scores for the Moviescope and MM-IMDb datasets. The BPMulT has also been tested on various multimodal datasets and has achieved competitive results in the IEMOCAP and CMU-MOSEI datasets.

Another marvelous thing is that the BPMulT has been tested even when we have less than three modalities, like in the MM-IMDb dataset. The proposed extension to handle this kind of dataset is that the BPMulT takes the sequence of one modality with different embedding representations.

9. Future Work

The BPMulT considers the single projections and the biprojections using residual connections and FGMU modules. With a TSNE study, we found that sometimes, a crossmodal transformer does not have an apparent label clustering. It happens with not the same modalities, projections, and biprojections and is different in each dataset. We believe that it could be a specific line to follow. To develop an architecture that could automatically detect whether one crossmodal transformer has a good clustering with the TSNE study or not. If the architecture detects that it does not have a good clustering, it will not have relevant information for the prediction.

Another line to follow is how to reduce the BPMulT because it is a heavy architecture. It could be addressed by taking an efficient crossmodal attention module with attention from two modalities, i.e., the biprojection packed in one attention module.

10. Acknowledgments

We would like to acknowledge CIMAT (Centro de Investigación en Matemáticas), INAOE (Instituto Nacional de Óptica y Electrónica) and CONACyT (Consejo Nacional de Ciencia y Tecnología) for the support to perform my research.

CRediT authorship contribution statement

Diego Aarón Moreno-Galván: Conceptualization of this study, Experiments, Proposals, Methodology, Software, Writing, Paper Draft. **Adrián Pastor López-Monroy:** Conceptualization of this study, Experiments, Proposals, Guidance of methodology. **Luis Carlos González-Gurrola:** Conceptualization of this study, Experiments, Proposals, Guidance of methodology.

References

Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A., 2017. Gated multimodal units for information fusion. Workshop track - ICLR URL: https://arxiv.org/pdf/1702.01992.pdf.

Baltrušaitis, T., Ahuja, C., Morency, L.P., 2017. Multimodal machine learning: A survey and taxonomy. Arxiv URL: https://arxiv.org/pdf/1705.09406.

Cascante-Bonilla, P., Sitaraman, K., Luo, M., Ordonez, V., 2019. Moviescope: Large-scale analysis of movies using multiple modalities. ArXiv abs/1908.03180.

Chandrasekaran, G., Nguyen, T., Hemanth, J., 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery URL: e1415.

Charland, P., Léger, P., Sénécal, S., Courtemanche, F., 2015. Assessing the multiple dimensions of engagement to characterize learning: A neurophysiological perspective. JoVE URL: doi:10.3791/52627.

Chauhan, P., Sharma, N., Sikka, G., 2021. Multimodal Sentiment Analysis of Social Media Data: A Review. pp. 545–561. doi:10.1007/978-981-15-8297-4_44.

Dai, W., Cahyawijaya, S., Liu, Z., Fung, P., 2021. Multimodal end-to-end sparse model for emotion recognition. Proceedings of the 2021 Conference of the North American , 5305–5316URL: https://arxiv.org/pdf/2103.09666v3.pdf.

Dai, W., Liu, Z., Yu, T., Fung, P., 2020. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. Proceedings of the 1st Conference of the Asia-Pacific , 269–280URL: https://arxiv.org/pdf/2009.09629.pdf.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. Arxiv URL: https://arxiv.org/pdf/1512.03385.

Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D., 2019. Supervised multimodal bitransformers for classifying images and text. Arxiv URL: https://arxiv.org/abs/1909.02950.

Nikolić, M., Majdandžić, M., Colonnesi, C., de Vente, W., Möller, E., Bögels, S., 2020. The unique contribution of blushing to the development of social anxiety disorder symptoms: results from a longitudinal study. J. Child Psychology Psychiatry URL: doi:10.1111/jcpp.13221.

Rodríguez-Bribiesca, I., López-Monroy, A.P., y Gómez, M.M., 2021. Multimodal weighted fusion of transformers for movie genre classification. Proceedings of the Third Workshop on Multimodal Artificial Intelligence , 1–5URL: https://aclanthology.org/2021.maiworkshop-1.1.pdf.

Sleeman-IV, W., Kapoor, R., Ghosh, P., 2021. Multimodal classification: Current landscape, taxonomy and future directions. Arxiv URL: https://arxiv.org/pdf/2109.09020.pdf.

Sourav, S., Ouyang, J., 2021. Lightweight models for multimodal sequential data. Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 2021, 129–137.

Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R., 2019. Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics , 6558–6569URL: https://aclanthology.org/P19-1656.pdf.

Vaswani, A., Jones, L., Shazeer, N., Parmar, N., Uszkoreit, J., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. 31st Conference on Neural Information Processing Systems URL: https://arxiv.org/pdf/1706.03762.pdf.

Xu, P., Zhu, X., Clifton, D.A., 2022. Multimodal learning with transformers: A survey. Arxiv URL: https://arxiv.org/pdf/2206.06488.

Yao, Y., Papakostas, M., Burzo, M., Abouelenien, M., Mihalcea, R., 2021. Muser: Multimodal stress detection using emotion recognition as an auxiliary task. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021, 2714–2725.

Author biography without author photo. Author biography.

Author biography with author photo. Author biography.

Author biography with author photo. Author biography.