# Multimodal Weighted Fusion of Transformers for Movie Genre Classification

**Isaac Rodríguez-Bribiesca** and **A. Pastor López-Monroy**
Mathematics Research Center (CIMAT)
GTO, Mexico
{isaac.bribiesca,pastor.lopez}@cimat.mx

**Manuel Montes-y-Gómez**
National Institue of Astrophysics, Optics and Electronics (INAOE)
Puebla, Mexico
mmontesg@inaoep.mx

## Abstract

The Multimodal Transformer showed to be a competitive model for multimodal tasks involving textual, visual and audio signals. However, as more modalities are involved, its late fusion by concatenation starts to have a negative impact on the model's performance. Besides, interpreting model's predictions becomes difficult, as one would have to look at the different attention activation matrices. In order to overcome these shortcomings, we propose to perform late fusion by adding a GMU module, which effectively allows the model to weight modalities at instance level, improving its performance while providing a better interpretabilty mechanism. In the experiments, we compare our proposed model (MulT-GMU) against the original implementation (MulT-Concat) and a SOTA model tested in a movie genre classification dataset. Our approach, MulT-GMU, outperforms both, MulT-Concat and previous SOTA model.

## 1 Introduction

Information on the internet has grown exponentially. Much of this information is multimodal (e.g. images, text, videos, etc.). For example, in platforms like YouTube and Facebook, multiple modalities can be extracted like video frames, audio and captions on different languages. In this context, it becomes increasingly important to design new methods that are able to analyze and understand automatically these type of multimodal content. One popular scenario is the movie streaming service (e.g. Netflix, Prime Video, etc.), where there is also an increasing interest in performing automatic movie understanding. In this paper we take as a case study the task of movie genre prediction. Our proposal exploits movie trailer frames and audio, plot, poster and a variety of metadata information, via Deep Learning techniques that have enough flexibility to fuse and learn to weight from all these modalities in a simultaneous way.

The success of the Transformer architecture (Vaswani et al., 2017) and its variants in NLP, has also inspired researchers to propose and extend these architectures in multimodal settings. Some examples include ViLBERT (Lu et al., 2019), MulT (Tsai et al., 2019), VisualBERT (Li et al., 2019), UNITER (Chen et al., 2020), MMBT (Kiela et al., 2019) and LXMERT (Tan and Bansal, 2019). However, the vast majority of these multimodal architectures were designed and tested only on bimodal data, more specifically, on text and visual information. Besides, models that only allow for early fusion, have the disadvantage that they rely solely on this mechanism that hinders the interpretability. While in models that output a feature per modality, an additional late fusion mechanism can be implemented to further fuse modalities and learn a richer representation, which is the case for the MulT model. Nonetheless, late fusion in this model was originally performed by means of concatenation, diminishing its fusion capacity.

Contributions of this work are twofold: We first adapt the MulT model (Tsai et al., 2019) to support additional number of modalities. Then, we consider a mechanism that learns to fuse all the modalities dynamically before making the prediction over each particular instance. This is a crucial step, given that for movies belonging to different genres, the relevant modalities could be quite different. For example, in Animation movies, visual information might be more relevant given the visual style, while for Drama movies, sound may be more helpful because of loud noises and screams.

In order learn to fuse the final representation of each modality we propose to adapt the GMU module (Arevalo et al., 2019). These units are highly interpretable gates which decide how each modal-



*Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 1–5
June 6, 2021. 

ity influences the layer output activation units, and therefore, decide how relevant each modality is in order to make the prediction. This is a crucial step, given that for this task, not all modalities are going to be equally relevant for each observation, as has been shown is previous work like (Mangolin et al., 2020) and (Cascante-Bonilla et al., 2019). Our evaluation shows that our MulT-GMU model, which uses weighted fusion by GMU, can outperform SOTA results in the movie genre classification task by 4%-10% on all metrics ($\mu$AP, mAP and sAP).

We explore for the first time the use of the MulT in the movie genre prediction task. We demonstrate that the original MulT model, which uses late fusion by concatenation (MulT-Concat) can achieve SOTA results task for this . Then, we show that further improvements can be achieved by our proposed model with the GMU module (MulT-GMU). The contributions can be summarized as follows:

- We introduce the use of the Multimodal Transformer architecture (MulT) to the task of movie genre prediction.
- We improve the MulT model by including a GMU module on the top, which allows to successfully fuse more modalities and improve its prediction performance.
- We show that the interpretability of the MulT model increases by incorporating the GMU module, allowing to better understand the relevance of each modality for each instance.

## 2 Approach

In Sections 2.1 and 2.2, we briefly describe the MulT architecture and then explain how to adapt the GMU units at the top of the model to perform a more robust late fusion of modalities.

### 2.1 Multimodal Transformer (MulT)

In (Tsai et al., 2019) the MulT model was proposed in the context of human multimodal language understanding, involving a mixture of natural language, facial gestures, and acoustic behaviors. Thus, it operates with three different modalities, Language (L), Video (V) and Audio (A).

Each modality is represented as a sequence of features $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ with $\alpha \in \{L, V, A\}$ being the modality. $T_{(.)}$ and $d_{(.)}$ are used to represent sequence length and feature dimension, respectively. Sequences are fused by pairs through crossmodal attention modules. These modules take two input modalities, $\alpha, \beta \in \{L, V, A\}$, and their respective sequences, $X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$ and $X_\beta \in \mathbb{R}^{T_\beta \times d_\beta}$. The crossmodal attention block will try to adapt latently the modality $\beta$ into $\alpha$. To achieve this, queries from one modality are combined with keys and values from the other modality. $D$ crossmodal transformer layers are stacked to form a crossmodal transformer. Another crossmodal transformer is used to provide the latent adaptation of modality $\alpha$ into $\beta$. Yielding representations $Z_{\beta\to\alpha}$ and $Z_{\alpha\to\beta}$, respectively.

In the case of three modalities (L, V, A), six crossmodal transformers are needed in order to model all pair interactions. Interactions that share the same target modality are concatenated. For example, the final representation of Language will be $Z_L = [Z^{[D]}_{V\to L}, Z^{[D]}_{A\to L}] \in \mathbb{R}^{T_{\{L,V,A\}} \times 2d}$. Finally, each modality is passed through $L$ transformer encoder layers, separately. The last element of each sequence is concatenated and passed through fully connected layers to make predictions.

### 2.2 MulT-GMU: Extending MulT through GMU-based late fusion

The MulT model expects the inputs to be sequences of features, but there could be modalities that are not sequences but a fixed vector (e.g. an image). A simple approach would be to concatenate them alongside the MulT outputs ($Z_L, Z_V, Z_A$) just before the fully connected layers. We argue that this is not optimal given that the fully connected layers will not be able to properly weight the relevance of each modality. In this work, we propose to adapt the MulT model by changing the concatenation fusion with a GMU module, as shown in Figure 1.

The GMU module receives a feature vector $x_i \in \mathbb{R}^{d_i}$ associated to modality $i$. Then the associated gate, $z_i \in \mathbb{R}^{shared}$, controls the contribution of that modality to the overall output of the GMU module. For this, the first step is to calculate an intermediate representation, $h_i = tanh(W_i x_i^T) \in \mathbb{R}^{shared}$ with $W_i \in \mathbb{R}^{shared \times d_i}$, where all modalities have the same dimension so they can be added and weighted by $z_i$. The next step is to calculate the gates $z_i = \sigma(W_{z_i}[x_i]^N_{i=1}) \in \mathbb{R}^{shared}$ where $N$ is the number of modalities and $[x_i]^N_{i=1}$ means the concatenation of vectors from $x_1$ to $x_n$. Finally, given the gates $z_1, z_2, ..., z_N$ and hidden features $h_1, h_2, ..., h_N$, fusion is performed through $h = \sum_{i=1}^{n} z_i \odot h_i$, where $\odot$ represents component-wise vector multiplication. This operation allows

the GMU module to have a global view of all modalities, whereas MulT only allows for early fusion by modality pairs.

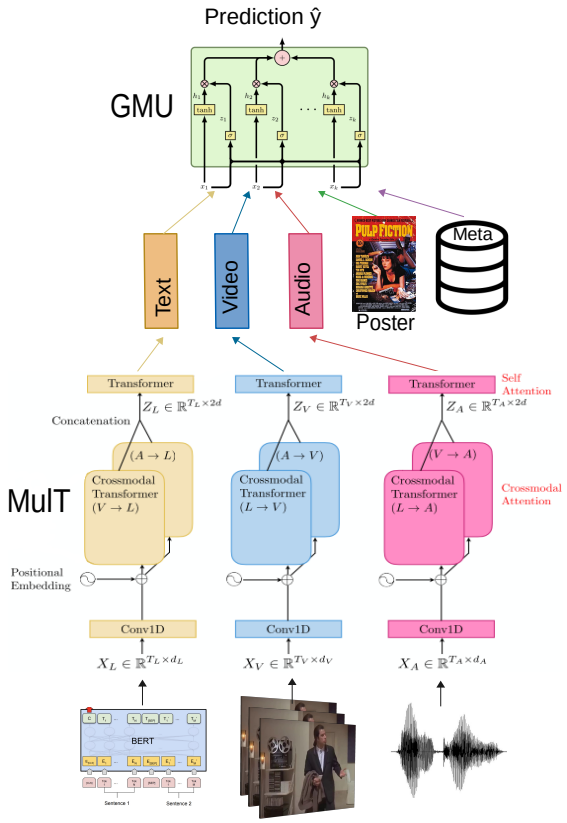


Figure 1: Proposed extension to the MulT architecture with a GMU module.

## 3 Evaluation

### 3.1 Datset

We base all of our experiments in the dataset Moviescope (Cascante-Bonilla et al., 2019). This is a large-scale datset comprising around 5,000 movies with corresponding movie trailers (video and audio), movie posters (images), movie plots (text), and metadata. The available data is already preprocessed. For the trailer video, we have 200 features vectors of size 4096, associated to 200 video frames subsampled by taking 1 every 10 frames. For the audio, log-mel scaled power spectrograms are provided. Poster images are provided in both, raw format and as a feature vector of size 4096. For the plot and metadata, raw data is provided. In the case of text, we use the pre-trained BERT-base model to extract features. For the metadata we follow (Cascante-Bonilla et al., 2019), extracting 13 different metadata values concatenated as a vector.

### 3.2 Experimental Framework

We compare three different models. The MulT model that works by concatenation of modalities (MulT-Concat), the extension of the MulT model with the GMU module (MulT-GMU), and the baseline model proposed in (Cascante-Bonilla et al., 2019), which is inspired by fastText (Joulin et al., 2017) to encode a sequence of features from text into a single vector, and a sequence of video features extracted from a pre-trained CNN also into single vector. The fusion of modalities is performed through a weighted regression, which could be considered as a form of modal attention. We refer to this model as Fast Modal Attention (Fast-MA).

In the case of the MulT-Concat and MulT-GMU, we show their mean performance over 5 runs with different random seeds. For the Fast-MA model we include the original results presented in (Cascante-Bonilla et al., 2019). The different modalities are denoted as V (Video), A (Audio), P (Poster), T (Text) and M (Metadata). The Fast-MA model was only tested in four of the presented settings (VA, VAP, TVAP and TVAPM). Furthermore, to investigate the impact of the GMU module we also include a more exhaustive list of experiments.

### 3.3 Results

We compared both baseline models, Fast-MA, MulT-Concat (late fusion by concatenation) with our proposed architecture MulT-GMU. Results on four different modality settings are shown in Table 1. They indicate that both MulT-Concat and MulT-GMU were able to outperform the state-of-the-art model Fast-MA when several modalities are considered. These results also show that Fast-MA outperformed both MulT-Concat and MulT-GMU in two of the modality settings, namely VA (Video and Audio) and VAP (Video, Audio and Poster). Note that these two settings are the only ones where Text (T) is not included, which confirms previous studies showing that for this task, text is the most relevant modality while audio is the least relevant (Mangolin et al. (2020), Cascante-Bonilla et al. (2019)). This explains in part, the low performance of the MulT models in these two settings. Once text is included, performance in MulT models increases dramatically. For example, from Table 2, we show that either bimodal MulT model that included text (TV or TA) already outperformed the best Fast-MA model (TVAPM).

Once we show the outstanding performance of

both MulT models, in Table 2 we further compare them on more modality settings. We can see that MulT-GMU outperforms MulT-Concat in almost all the settings except in TV (Text and Video). For example, from experimental settings TVPM and TVAPM, we can observe that MulT-Concat has difficulty handling the Metadata features, dropping quite considerably the performance. In contrast, MulT-GMU is able to handle these features and maintain or even increase its performance.

| Modality | Model | $\mu AP$ | $mAP$ | $sAP$ |
|---|---|---|---|---|
| VA | Fast-MA | **70.3** | **61.5** | **78.8** |
| | MulT-Concat | 59.2±0.3 | 53.1±0.5 | 71.1±0.7 |
| | MulT-GMU | 58.9±0.7 | 52.5±0.6 | 70.6±0.6 |
| VAP | Fast-MA | **70.4** | **61.7** | **78.8** |
| | MulT-Concat | 63.1±0.5 | 54.3±0.5 | 73.9±0.5 |
| | MulT-GMU | 64.1±0.9 | 55.0±0.7 | 74.5±0.5 |
| TVAP | Fast-MA | 74.9 | 67.5 | 82.3 |
| | MulT-Concat | 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** |
| TVAPM | Fast-MA | 75.3 | 68.6 | 82.5 |
| | MulT-Concat | 64.8±5.8 | 61.3±7.2 | 76.9±4 |
| | MulT-GMU | **79.5±0.5** | **76.4±0.3** | **85.6±0.3** |

Table 1: Comparison against MulT-Concat (Tsai et al., 2019) and Fast-MA (Cascante-Bonilla et al., 2019) on different modality combinations. Metrics reported correspond to average precision, micro ($\mu AP$), macro ($mAP$) and sample ($sAP$) averaged.

| Modality | Model | $\mu AP$ | $mAP$ | $sAP$ |
|---|---|---|---|---|
| TV | MulT-Concat | **77.5±0.5** | **73.5±0.2** | **84.4±0.2** |
| | MulT-GMU | 76.9±0.3 | 73.2±0.2 | 84.2±0.4 |
| TA | MulT-Concat | 76.2±0.7 | **72.4±0.8** | 84±0.5 |
| | MulT-GMU | **76.3±0.4** | 71.1±0.4 | **84.1±0.2** |
| TVA | MulT-Concat | 77.2±0.7 | 74.8±0.4 | 84.2±0.5 |
| | MulT-GMU | **78.2±0.5** | **74.9±0.5** | **85±0.3** |
| TVP | MulT-Concat | 78.4±0.5 | 75.1±0.4 | 85.1±0.5 |
| | MulT-GMU | **78.9±0.1** | **75.2±0.4** | **85.7±0.3** |
| TVPM | MulT-Concat | 46.1±11 | 43.2±10.7 | 62.8±8.8 |
| | MulT-GMU | **79.1±0.3** | **75.4±0.2** | **85.4±0.4** |

Table 2: Comparison of the proposed model MulT-GMU and MulT-Concat (Tsai et al., 2019) with additional modality combinations. Metrics reported correspond to average precision, micro ($\mu AP$), macro ($mAP$) and sample ($sAP$) averaged.

## 4 Qualitative analysis

To understand how the GMU units are weighting the relevance of each modality according to each instance (movie) $i$, we inspected the gates $z_i$ of the GMU module for all the observations in the test set. To achieve this, we selected the observations that contained each of the genres and averaged the gate activations per modality. We show results for 5 different movie genres in Figure 2, where each row already takes into account the average of all test movies of the corresponding genre.

In general, text and visual modalities were the most relevant according to the GMU module. We can see relatively low activations for the audio modality compared with the other ones. This is expected as it has been shown that audio modality is not as useful as the other ones, for this task (Mangolin et al. (2020), Cascante-Bonilla et al. (2019)). There is also a relationship between audio and video signals. In genres where video is the strongest, audio is the weakest.

Taking the Audio modality as an example, where Horror and Drama had the highest GMU activations overall, we could think that this was the case given that this kind of movies usually have loud noises like screams in the trailers, so this could be a good indicator that the movie is likely to belong to one of these two genres. There are other interesting scenarios, for example the text modality had the highest activation for genres like Comedy and Drama. In the case of the video modality, Comedy and Family genres had the highest activation.

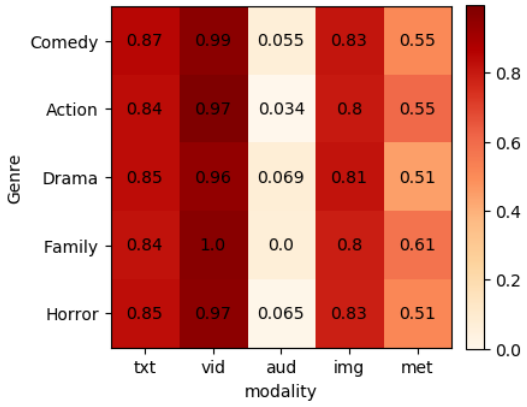


Figure 2: Average proportion of GMU unit activations normalized by genre for all the observations in test set. We only show the activations for 5 movie genres.

## 5 Conclusion

We proposed an adapted version of the Multimodal Transformer, MulT-GMU, by performing weighted late fusion with a GMU module. This approach achieved SOTA results in the multimodal movie

genre classification task. Moreover, we improved the interpretability of the MulT model by performing a qualitative analysis, visualizing the activations of the GMU module, which allowed us to have a better understanding about relevant modalities for the model, depending on the genre of the movie. To the best of our knowledge, this is the first time multimodal transformer-based architectures are tested in the task of movie genre classification.

## Acknowledgements

The authors thank CONACYT, INAOE and CIMAT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajio Supercomputing Laboratory (#300832). Rodríguez-Bribiesca would like to thank CONACYT for its support through scholarship #952419.

## References

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2019. Gated multimodal networks. *Neural Computing and Applications*, 32(14):10209–10228.

Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. 2019. Moviescope: Large-scale Analysis of Movies using Multiple Modalities.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. (2):1–14.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. pages 1–11.

Rafael B. Mangolin, Rodolfo M. Pereira, Alceu S. Britto, Carlos N. Silla, Valéria D. Feltrim, Diego Bertolini, and Yandre M. G. Costa. 2020. A multimodal approach for multi-label movie genre classification. pages 1–21.

Hao Tan and Mohit Bansal. 2019. LXMert: Learning cross-modality encoder representations from transformers. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5100–5111.

Yao Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 6558–6569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.