Deep Learning for Multimodal Classification

Diego Moreno

Technological Project, Department of Computer Science, Centro de Investigación en Matemáticas, Guanajuato, MX.
Email: diego.moreno@cimat.mx

In this technological project, we have reviewed recent research, datasets, and neural network architecture related to multimodal classification. Two lines of research were found interesting for us, Emotion Recognition and Movie Genre Classification task. For the Emotion Recognition task, we have gotten two datasets (CMU-MOSEI and IEMOCAP) and used CMU-MOSEI in an experiment. On the other hand, we downloaded and cleaned the Moviescope (complete) dataset for the Movie Genre Classification task and evaluated an experiment. Also, we have proposed a novel neural network architecture for the fusion of multimodal features. Experimental results suggest that the proposed architecture is better for movie classification. Further research on experiments has to be done.

Keywords: Multimodal Classification; Emotion detection; Movie Genre Classification; Hybrid Early-Late Fusion

Received 30 July 2021

1. INTRODUCTION

Multimodal classification has been an emerging research field at the intersection of natural language processing, computer vision, and speech recognition since the world's tremendous growth of social media and content platforms. Hence, the research on multimodal information processing has attracted the attention of researchers and developers. For example, a movie is a multimodal input that provides visual, acoustic, and textual information. The motivation for multimodal movie genre classification and emotion analysis lies in leveraging the variety of information from multiple sources to build more efficient systems. Sometimes, text can provide a better clue for the prediction, whereas, for others, acoustic or visual sources can be more informative. For example, only the text "you have been working hard" cannot decide if there is sarcasm, but acoustic (tone of a person) and visual (expression of a person) can reveal details about that. However, effectively fusing this diverse information is a non-trivial task that researchers often need help with.

In this technological project, we have been looking for recent research about the effective combining of multimodal information. Traditionally, "text" has been the critical factor in any Natural Language Processing tasks, such as emotion detection and movie genre classification. We have been interested in how researchers handle acoustic, text, and visual information to give a prediction of an emotion or a genre in a movie's context.

2. EMOTION ANALYSIS AND MOVIE GENRE CLASSIFICATION TASK

There are many ways to get multimodal information, such as social media or digital content. For the Emotion detection task, the interest is in human-computer interaction, an essential aspect of enhancing interpersonal relationships. Commonly, emotions could be happy, sad, angry, fearful, disgusted, and surprised. Here, there are too many aspects to deal with. For example, [1] discuss relevant data is hard to come by and notably costly to annotate, which poses a challenging barrier to building robust multimodal emotion recognition systems. Models trained on relatively small datasets tend to overfit, and the improvement gained by using complex state-of-the-art models is marginal compared to simple baselines. They proposed a novel way to deal with a small dataset and avoid overfitting.

Another research line explored by [2] is to find a relation between emotions. They propose a novel intuitive space representation of emotions, which can model these "relations" by closeness in the emotion space.

Similarly, [3] tries to give an approach for label-dependence modeling. In this paper, we have found a transformer-based architecture to model this dependence between emotions. Moreover, [4] research about context-aware attention for emotions, and they proposed a GRU-based architecture to model relations between information given by text, video, and audio.

Based on new ways to combine multimodal information for sentiments, [5] talks about an efficient way to do this, comparing it with three transformer-based architectures. These three correspond to fuse the video, audio, and text features early, late, or in a hybrid manner. It suggests that a hybrid combination of features improves emotion detection by remembering original data. Here is where we have seen a possibility of doing an experiment looking for an improvement on different tasks.

We have noticed that almost all of these multimodal-fusing architectures are based on Transformers. For example, for text features, most papers use a self-supervised bidirectional transformer model such as BERT since it has led to improvements in various text classification tasks. Also, there are strong baselines such as [6], which discuss a supervised multimodal bitransformer model that fuses information from text and image encoders and obtain state-of-the-art performance on its time.

The most crucial approach we have found is the Multimodal Transformer model was proposed in the context of human multimodal language understanding, involving a mixture of natural language, facial gestures, and acoustic behaviors. It works with three different modalities, Language (L), Video (V), and Audio (A). Projecting each modality to the space of another modality, a Bi-Transformer will work with it to get relevant features. This architecture has been used in several research papers for emotion recognition and recently in [7], which has been used for classifying movie genres.

3. DATASETS

In searching for papers related to Multimodal-fusing architectures, two standard datasets on emotion detection tasks were IEMOCAP and CMU-MOSEI, described below.

IEMOCAP, [8]. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) is a multimodal emotion recognition dataset containing 151 videos along with the corresponding transcripts and audio. In each video, two professional actors conduct dyadic dialogues in English. Multiple annotators annotate the IEMOCAP database into categorical labels, such as anger, happiness, sadness, and neutrality, and dimensional labels, such as valence, activation, and dominance.

CMU-MOSEI, [9]. The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is a multimodal emotion recognition and sentiment analysis dataset. It comprises 3,837 videos from 1,000 diverse speakers and is annotated with six emotion categories: happy, sad, angry, fearful, disgusted, and surprised. In addition, each data sample is also annotated with a sentiment score on a Likert scale [-3, 3].

For our experiment on movie genre classification, we use the Moviescope dataset, which is described below:

Moviescope, [10]. It is a large-scale dataset 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 feature vectors of size 4096, associated with 200 video frames subsampled by taking one every ten frames. For the audio, log-mel scaled power spectrograms are provided. Poster images are provided in raw format and as a feature vector of size 4096. For the plot and metadata, raw data is provided.

4. METHODS

We have coded on Python3 our adaption functions on and used the INAOE and CIMAT Bajío Supercomputing Laboratories. For an interactive and easy running, we recommend INAOE Lab, but for an efficient running is better to use CIMAT Lab. It is worth mentioning that it also takes a considerable amount of time to learn how these super-computer clusters work.

4.1. Emotion Detection Experiment

For this task and a simple running, we have selected the "Modality-transferable emotion embeddings for low-resource multimodal emotion recognition," [2], paper. This paper uses a multimodal transformer to perform emotion detection.

We have used the proportioned code by authors and the CMU-MOSEI dataset, which is already aligned to run an initial experiment.

The following parameters have been considered: With a batch size of 64, learning rate of 1e-3 during 100 epochs, a dropout probability of 0.15, a data sequence length of 20, and patience of non-improvement of 10.

Tables 1 and 2 show the test set is running results. We have gotten comparable and similar results that the author presents in their paper.

4.2. Movie Genre Classification Experiment

As we have seen, multimodal Transformers is one of the most famous architectures to fuse features given by audio, images, and text. For this experiment, we have been based on [7], which contains an architecture based on a multimodal transformer, and then it fuses each projection feature with a GMU. The first approach is to run the author's code and have similar results. After that, we propose a novel architecture that takes this architecture that we have mentioned before, and we add a hybrid fusion method to get a prediction of a movie genre.

An overall architecture view of the proposed model can be seen in Figure 1. The part on grayscale corresponds to our contribution to the given architecture. Part in color corresponds to the exact architecture on [7]. For this experiment, we have used training parameters: an initial learning rate of 1e-4, patience for non-improvement of 5 on epochs, and 3 for

Metric	Anger	Disgust	Fear	Happy
F1-score	0.687	0.729	0.752	0.678
AUC	0.696	0.776	0.688	0.732

TABLE 1. Part 1: Test results of Modality-transferable emotion embeddings for low-resource multimodal emotion recognition

Metric	Sad	Surprise	Average
F1-score	0.627	0.629	0.684
AUC	0.646	0.65	0.698

TABLE 2. Part 2: Test results of Modality-transferable emotion embeddings for low-resource multimodal emotion recognition

learning rate with a factor of 0.5, a maximum of 50 epochs.

To compare our results with the original paper as a baseline, we also performed the original architecture given by the authors with the complete dataset. We have trained the text with the pre-trained BERT embedding for text, also the audio and frames' part. The trained features will be used for both architectures (proposed and original) and trained over them.

Note that in the code, we only have text, video, and audio features; authors do not use the poster and metadata in the version they distribute (or I have). Hence, we got results training the networks (original and proposed) on our own.

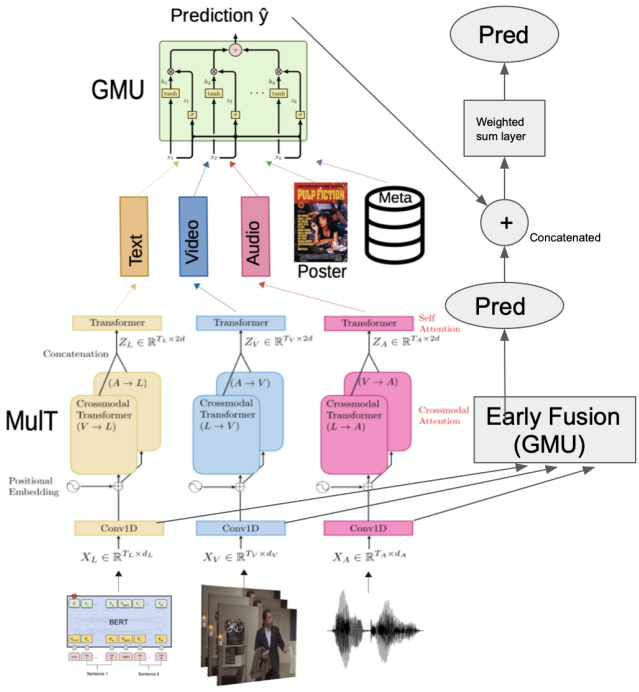

FIGURE 1. Overall architecture of the proposed model with hybrid fused features for multimodal classification.

5. RESULTS

5.1. Emotion Detection Experiment

Tables 1 and 2 show the running results for the test set. We have gotten comparable and similar results that the author presents in their paper.

Metric	Original Arch.	Proposed Arch.
AUC μP	0.524	0.560
AUC mP	0.465	0.582
AUC sP	0.654	0.685

TABLE 3. Comparison between original architecture [7] and proposed (Hybrid fused features) on text, audio, and video. Metrics reported corresponding to AUC of average precision, micro (μP), macro (mP), and sample (sP) averaged.

5.2. Movie Genre Classification Experiment

Table 3 compares our proposed improvement for the original architecture evaluated on the test set. Note that we have gotten comparable and better results on the complete dataset of Moviescope. However, these results are just with video, audio, and text, which are different from the author's results on its paper since they use, in addition, the movie's poster and metadata.

6. CONCLUSION

As a first conclusion, the project Deep Learning for Multimodal Classification has been an excellent opportunity to learn about current research topics. We have gained experience in research, running experiments on clusters, and have read a bunch of papers to get and know some ideas for handlining multimodal data.

On the other hand, we noticed that the Multimodal Transformer is one of the most used architectures for multimodal classification. It appears in emotion detection and movie genre classification as well. We researched how the authors are fusing the multimodal data, and we could detect emotion detection tasks, one good hybrid fusion that has yet to be implemented in movie genre classification. We have proposed a novel way to fuse data with a GMU on early fusion and a hybrid fusion to get a prediction. Experiments were done to suggest that this kind of multimodal fusion could significantly improve the Moviescope dataset. For further work, testing this new architecture with more datasets, the movie poster, and metadata is imperative.

ACKNOWLEDGEMENTS

Supported by CONACYT, INAOE, and CIMAT with computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies and CIMAT Bajio Supercomputing Laboratory. D. Morgal acknowledges CONA-

CYT for its support through scholarship.

REFERENCES

[1] Dai, W., Cahyawijaya, S., Bang, Y., and Fung, P. Weakly-supervised multi-task learning for multimodal affect recognition. *arXiv preprint arXiv:2104.11560*, **2021**, 3584–3593.

[2] Dai, W., Liu, Z., Yu, T., and Fung, P. Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, **2020**, 269–280.

[3] Zhang, D., Ju, X., Li, J., Li, S., Zhu, Q., and Zhou, G. Multi-modal multi-label emotion detection with modality and label dependence. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, **2020**, 3584–3593.

[4] Chauhan, D. S., Akhtar, M. S., Ekbal, A., and Bhattacharyya, P. Context-aware interactive attention for multi-modal sentiment and emotion analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, **2019**, 5647–5657.

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

[6] Kiela, D., Bhooshan, S., Firooz, H., Perez, E., and Testuggine, D. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, **2019**, 0–5.

[7] Rodríguez-Bribiesca, I., López-Monroy, A. P., and y Gómez, M. M. Multimodal weighted fusion of transformers for movie genre classification. *Proceedings of the Third Workshop on Multimodal Artificial Intelligence, Association for Computational Linguistics*, **2021**, 1–5.

[8] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Provost, E. M., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, **2008**, 0–5.

[9] Zadeh, A., Liang, P. P., Poria, S., Cam-bria, E., and Morency, L.-P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. *ACL*, **2018**, 0–5.

[10] Cascante-Bonilla, P., Sitaraman, K., Luo, M., and Ordonez, V. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180*, **2019**, 0–5.

[11] Yao, Y., Papakostas, M., Burzo, M., Abouelenien, M., and Mihalcea, R. 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.